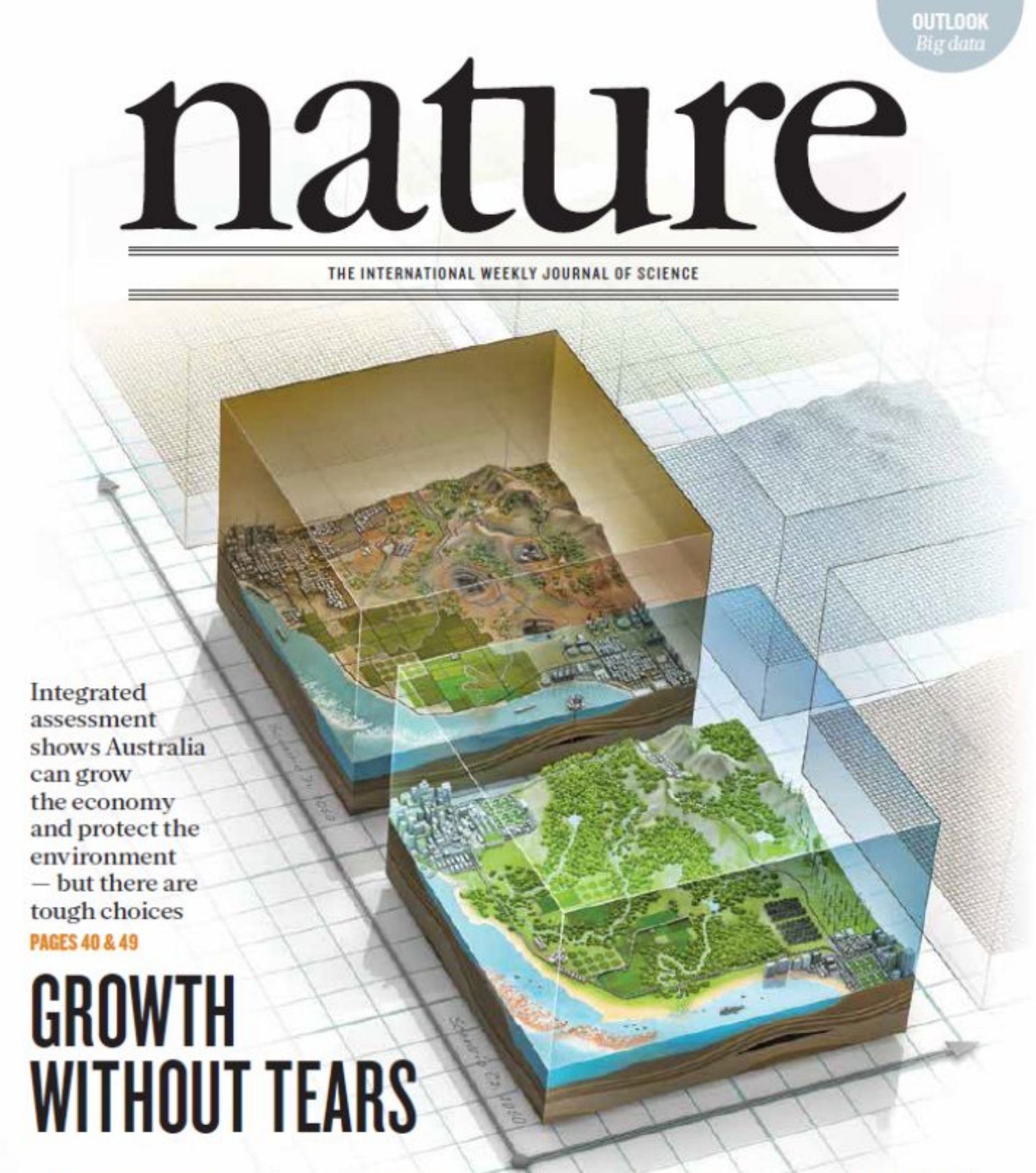
nature

THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE



Integrated assessment shows Australia can grow the economy and protect the environment — but there are tough choices
PAGES 40 & 49

GROWTH WITHOUT TEARS

AMERICAN PREHISTORY

MESA VERDE MYSTERY

Why Ancestral Puebloans left their cosy caves

PAGE 26

SOCIETY

HEALTHY DATA USE

Leverage personal information for all

PAGES 31 & 33

SOFTWARE

KNOWLEDGE YOU CAN USE

The joy of digital reference management

PAGE 123

NATURE.COM/NATURE
5 November 2015 £10
Vol. 527, No. 7576

THIS WEEK

EDITORIALS

LOGIC The digital legacy of George Boole's smooth operators **p.8**

WORLD VIEW Politicians should let problems slip and slide **p.9**

TECHNOLOGY Magnets tune and tweak the 3D printing process **p.11**

Ills of the system

Reform is long overdue for Germany's archaic medical-education system, which puts undue pressure on students and contaminates the scientific literature.

Of all the problems on the desk of the German defence minister Ursula von der Leyen, accusations of plagiarism in her quarter-century-old medical thesis may not seem to rank very highly. Yet similar allegations claimed the scalp of her predecessor.

Although plagiarism is a universal plague in academia, Germany has its own distinct circumstances. Almost uniquely among nations, most German medical students must squeeze out a doctoral thesis during their years of full-time training. Many of these theses, not surprisingly, are not very good. Corners are cut and quality suffers.

The high-profile case of von der Leyen's 1990 dissertation, first publicized in September by the web platform VroniPlag Wiki, which searches theses for plagiarism, should bring change — but not in the government. It is Germany's antiquated medical-education system that must be reformed.

Von der Leyen — who denies misconduct charges and has asked Hanover Medical School, where she studied, to investigate — is hardly alone. Evidence that the system of medical doctorates is failing has been accumulating for years.

Thousands of these dissertations are produced every year in Germany and plagiarism is far from the only problem. The DFG, Germany's research agency, and the Wissenschaftsrat, its high-level science council, have over the years drawn attention to more fundamental problems, such as study design and analysis. Some experts privately say that most medical theses are scientifically valueless.

Germany justifiably takes pride in its long tradition, and high standards, in science. So what is going so badly wrong in its medical faculties? In most countries, medical students receive their medical degree — and 'Dr' title — after successfully completing both preclinical undergraduate studies and clinical training, and then passing a state examination. Not so in Germany, where the degree gives them only the right to practise medicine — not to title themselves Dr. To acquire that honour, an extra step is required: a research project leading to a thesis, done, written up and published in the student's spare time. Most students choose to do this: after all, what ill person wants to visit a doctor who does not bear that title? But in the busy, frequently self-important, world of the clinical sciences, supervision is often inadequate.

In 2004, the Wissenschaftsrat called for an end to this system and the laxness that it actively encourages. It recommended that medical students get their medical degree and doctor title automatically, without having to do a research thesis. Students with genuine interest in medical science, it said, should have the option of taking time out to do a PhD to the same standards as other sciences.

Because the Wissenschaftsrat includes representatives of federal and state governments as well as top scientists, its recommendations are usually implemented. But the call for the automatic degree and title — which would require a change of federal law — fell on deaf ears. Medical faculties ignored it, although many have established graduate schools to make available an alternative route to a high-quality PhD.

The value of those recommendations has not changed in the past decade, however. Good graduate colleges for the medical sciences are fundamental to the drive to speed basic-research discoveries into the clinic, an ambition that requires research-savvy physicians. But it makes no sense to maintain the requirement for a quick-and-dirty thesis, which adds stress to medical students who are already under immense pressure, while teaching them little beyond the dangerous lesson that it is acceptable for medical science to be sloppy.

"It makes no sense to maintain the requirement for a quick-and-dirty thesis."

In 2010, the DFG published a strongly worded report calling for scientific standards in medical dissertations to be raised, and earlier this year the German Rector's Conference (HRK) established a task force to look into the problem. However, like the DFG and the Wissenschaftsrat, the HRK will be able to do no more than make recommendations. Medical faculties and the profession in general now have to decisively shed their reluctance to abandon their aberrant doctoral system. They should do so, before the public shame becomes unbearable. How many medical theses need be exposed on VroniPlag Wiki — which already hosts dozens of examples, some quite brazen — before the bankruptcy of the system is accepted?

Plagiarism can never be defended. But the pressures on medical students — many of whom do not resort to plagiarism in response — make the temptation to indulge understandable. Von der Leyen may simply have been a student of her times — times that now have to change. ■

Care for the carers

Researchers should add their voices to the effort to stop attacks on health workers in war zones.

As the world this week commemorates the armistice that ended the First World War in 1918, it is reprehensible that humanitarian rules forged in the suffering and bloodshed of battle are often being violated in contemporary conflicts. In the past month alone, two hospitals run by Médecins Sans Frontières (MSF; also known as Doctors Without Borders) were hit by air strikes. US warplanes destroyed one in Kunduz in Afghanistan — killing 13 MSF staff and 17 others — and another in Yemen was targeted, allegedly by Saudi-led coalition forces.

These are not isolated incidents, but part of a string of violations of a fundamental part of international humanitarian law — that warring

parties must consider the wounded and the medical staff who care for them as neutral, and protect them from harm.

The public and the media must increase calls for political and diplomatic pressure to help to prevent such attacks. The scientific community, and in particular biomedical and clinical researchers and the professional bodies that represent them, must add their voices to this timely and important matter.

The need for ground rules in conflicts has been recognized since antiquity, but today's international humanitarian laws have their roots in the work of the nineteenth-century Swiss businessman, Henry Dunant. Horrified by the thousands of wounded left untreated and dying on the battlefield after the French and Sardinians crushed the Austrian army at Solferino in Italy in 1859, he proposed that states should allow, and protect, humanitarian volunteers to care for those who are wounded.

In 1863, he helped to found what was to become the International Committee of the Red Cross (ICRC). Dunant's efforts spurred 16 countries to agree the following year to the first internationally codified rules of war; the first Geneva Convention for the Amelioration of the Condition of the Wounded and Sick in Armed Forces in the Field. As well as granting neutral status to medical staff, it obliged warring parties to care for wounded enemy prisoners.

As the nature of warfare has changed, so the wording and scope of the Geneva Conventions have been regularly revised — for example in 1949 to better protect civilians. The principle of medical neutrality is more relevant today than ever, but it is under increasing threat.

Syria, where conflict sparked in 2011, is by far the worst case. As of the end of September, 313 attacks on 227 medical facilities had been reported — 283 of them carried out by government forces, often using indiscriminate 'barrel bombs' dropped from helicopters. Over the same period, 679 medical staff have been killed, almost all by government forces, and scores of others have been arrested, imprisoned or tortured. The regime has also deployed chemical weapons. The health system has been all but destroyed in large parts of the country.

During peaceful protests in Turkey in 2013 and 2014, the government used violence against clinics and medical staff, and health workers have been arrested and charged with assisting criminals simply for having treated wounded protestors. Similarly, during protests against the government in Bahrain in 2013, doctors and nurses were fired from civil-service posts, then arrested and jailed for the same motive as those in Turkey. Dozens of workers dispensing polio vaccinations have been assassinated in Pakistan and Nigeria. The ICRC has identified almost 2,000 incidents of violence against patients, health workers and medical facilities in 23 countries in 2012 and 2013 alone.

These are estimates, but comprehensive monitoring of violations and data are both lacking. However, Susannah Sirkin, director of international policy and partnerships for the humanitarian group Physicians for Human Rights, based in New York City, points out that "we can safely say that the bombing of hospitals and deliberate killing of hundreds of medics, especially in Syria, is something more extreme and extensive than we have ever seen".

Among the explanations is a lack of awareness of the Geneva Conventions by protagonists — in what are increasingly not wars between nations, but smaller civil and sectarian wars, often involving non-state actors — but also a poor grasp by the media and public. People may have "become inured to the extraordinary level of targeting of civilians in many conflicts in the past few decades and simply shrug at the inclusion of medical facilities as regular targets", adds Sirkin. What is worrying, she says, is that the overt targeting of humanitarian and health workers has become the "new normal", despite it being illegal under international law — and having the effect of depriving entire populations of health care, and children of vital vaccinations.

"The Geneva Conventions lack a body with teeth to ensure that the rules are respected."

But above all, abuses happen because there is little accountability, with perpetrators operating with almost total impunity, despite their actions often clearly amounting to war crimes — or indeed crimes against humanity. The Geneva Conventions lack a body with teeth to ensure that the rules are respected, or to stop abuses when they are under way. They also lack mechanisms to investigate and prosecute abuses.

Accountability has also suffered because many of those affected are voiceless. MSF, by contrast, has both political clout and moral authority, and, for example, is robustly and rightly pressing for an independent international fact-finding commission under the Geneva Conventions into the attacks on its facilities.

Momentum to stop the attacks, led by campaigns from humanitarian groups, is building within civil society. Meanwhile, Ban Ki-moon, the secretary-general of the United Nations, and Peter Maurer, the president of the ICRC, last week issued a joint warning about the unprecedented level of violations of international humanitarian law in ongoing conflicts.

As well as the armistice, this month marks 100 years since the decision to evacuate troops from the ill-fated 1915 Gallipoli campaign, in which medical staff working under atrocious battlefield conditions suffered extensive casualties. The world has been shocked into action to protect health workers before. It must be again. ■

Smooth operator

A tribute to the nineteenth-century polymath whose algebra lets you search the Internet.

IF George Boole had lived, then he would have celebrated his 200th birthday this week. NOT that it makes any sense to say such a thing OR to write it. People do NOT live that long. And if there was one thing that George Boole is known for, it is logic. AND mathematics AND philosophy. Three things. NOT one thing.

The combination of mathematics AND logic AND philosophy is NOT easy for many people to follow OR understand. So Boole is usually associated with the three words NOT AND OR. They are called Boolean operators AND they can be combined to make AND NOT. That's because the Boolean operator OR does NOT really mean OR, which usually means AND NOT.

It is NOT always easy to follow these logical constructions when they are written in words. That is why so many people call George Boole a genius. Because he did NOT have the same trouble. AND because he invented them OR applied them to mathematics. Without George Boole, people say that the modern world would NOT have been the same, with no computers OR electronics. Although a nice thing to say, it is probably NOT true AND someone else could have come up with the idea OR something similar. After all, Boole himself is a good example, who shows that ideas AND NOT inventions can come from an unlikely source.

He did NOT have a formal education OR academic training. He taught himself languages including Latin AND Greek AND calculus. He wrote scientific papers on how to represent logical relations as symbols AND algebraic equations. Despite NOT having a university education, he was appointed professor of mathematics at Queen's College Cork in Ireland.

The weather in Ireland is often NOT dry and Boole caught pneumonia after walking to the college in heavy rain. His wife, Mary, a prominent mathematician, was NOT as skilled at medicine. She soaked her husband's sheets with water AND made him shiver with cold. It did NOT help AND, sadly, he died. ■

➲ NATURE.COM
To comment online, click on Editorials at:
go.nature.com/xhunqv


LEN FISHER

Avoid major disasters by welcoming minor change

Scientists can educate policymakers on how to deal with the European refugee crisis — it's all about alleviating the pressure, says **Len Fisher**.

What can a 70-year-old book on how to play bridge tell us about addressing the ongoing refugee crisis in Europe? And what does it have to do with *King Lear*?

In Shakespeare's play, the Duke of Albany warns that "striving to better, oft we mar what's well". In the search for a solution, in other words, we can let the perfect become the enemy of the good. In his 1945 book *Why You Lose at Bridge*, S. J. Simon called it the half-loaf strategy: the most successful players aim for the best possible result, rather than the best result possible.

In human and political crises, the best possible result is often one of damage limitation — an outcome that avoids or delays the chance of a large-scale and catastrophic change. So, the question then becomes: how can we achieve such an outcome?

In a recent editorial (see *Nature* **525,** 157; 2015), *Nature* suggests that nations should "keep a welcome" for refugees. I agree. This pressure-releasing approach could serve as an effective paradigm for policy development, being used to handle emergent crises of many types. It makes sense: not just from a humanitarian perspective but also from what we now understand about the underlying behaviour of our interconnected global socio-economic-ecological system. For example, convincing connections have been drawn among the European refugee crisis, global warming and future food supplies, so a solution to one problem is likely to bear on the others.

Such complex systems can undergo sudden change at any time. These changes (known technically as 'regime shifts', 'critical transitions' or 'catastrophic bifurcations') occur at all scales, happen with little warning and often have no apparent cause. They frequently seem to be out of our control. Examples include cascading failure in power grids, communication networks, financial systems, food webs and social organizations; epidemics, not just of disease but also of social unrest and innovation; and sudden shifts in the balance of power, be they in international relationships or small groups.

Policy development to deal with such sudden change is often based on searching for (or blaming) specific causes. But, to quote H. G. Wells, "History is a race between education and catastrophe." Scientists must show policymakers that sudden change is inevitable in any complex system, and the first step towards avoiding or minimizing catastrophe is to recognize this.

The second step is to understand the nature of these transitions. Scientists have modelled them as, for example, sudden slippages in a sand pile when extra grains are progressively added, or as evolving interactions of multiple positive and negative feedback loops in a system. An important common feature of these models is that they predict that smaller changes are more frequent than larger ones.

WE SHOULD PROMOTE SMALLER, LESS-DAMAGING TRANSITIONS TO REDUCE THE CHANCE OF LARGER ONES.

Scientists have also suggested a number of different ways to develop policies to deal with the potential for sudden change. When it comes to protecting against terrorist attacks, some suggest that we must concentrate resources to protect critical nodes in a network. On the stability of banking systems, researchers argue for structural changes in networks so that damage in one part cannot easily propagate to others. A third idea, which to some extent complements the first two, is to build more resilience into our societies and institutions.

These ideas have their merits, but the 'keep a welcome' strategy for the refugee crisis suggests a different approach — one that can more easily be adapted to take account of the important (and sometimes overwhelming) human dimension in many crises.

This approach, which my colleague Jim Gimzewski and I have been examining, involves reducing the chances of sudden, large-scale, damaging change by altering the shape of the statistical distribution of event sizes. We should promote smaller, less-damaging transitions to reduce the chance of larger ones occurring. In metaphorical terms, the aim should be to reduce pressures before they can build to dangerous levels.

This is not a new principle. It underpins, for instance, the practice of triggering small snow avalanches to reduce the probability and impact of a major one, and it also has parallels with the philosopher Karl Popper's idea of 'piecemeal social engineering'. A simple social example is the reduction of traffic congestion by breaking the traffic into manageable blocks that are separated and accompanied by slow-moving police cars. Progress is still slow, but on average it is much faster than if large traffic jams were allowed to develop.

Most social problems, of course, are not quite this simple. But we must be wary of the 'nirvana effect' — the belief that perfect solutions are out there somewhere. A half-loaf of bread is always better than none at all. Thus, for example, in the case of the Greek debt crisis, our approach suggests paying to create jobs, rather than imposing austerity. The cost of the former would be far less than the social and economic costs that may result from the latter.

The current refugee crisis falls into a similar category. Countries willing to bear the (financial and political) cost of welcoming more refugees with fewer restrictions would promote small-scale changes that release the build-up of devastating social pressures. In this way, scientific and humanitarian values can work hand in hand. ■

➲ NATURE.COM
Discuss this article online at:
go.nature.com/tmgk72

Len Fisher *is a visiting research fellow at the University of Bristol, UK.*
e-mail: len.fisher@bristol.ac.uk

RESEARCH HIGHLIGHTS

Selections from the scientific literature

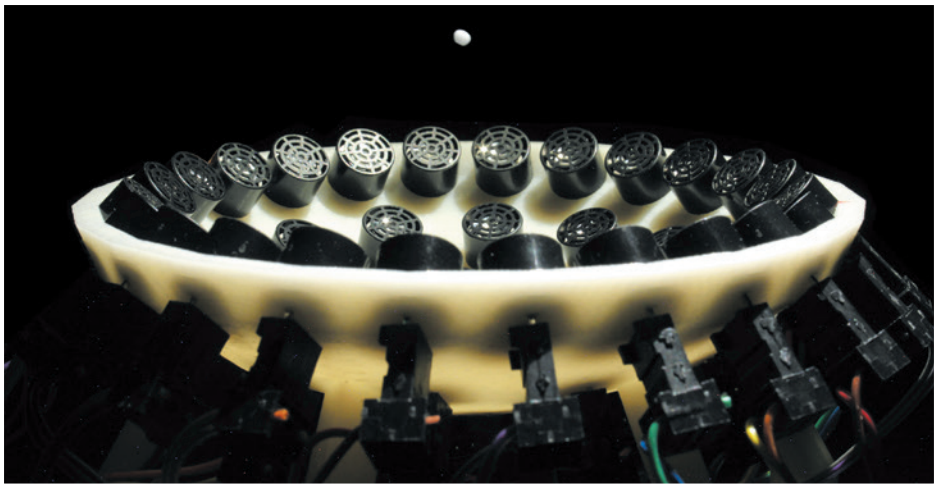
ECOLOGY

Carnivores curbed mammoth numbers

Sabre-toothed cats and other large carnivores were probably able to hunt down young mammoths and mastodons during the Pleistocene epoch, between 2.6 million years and 12,000 years ago. That would explain why Earth's forests were not grazed to death by the large numbers of big herbivores before they went extinct.

Researchers have long thought that mammoths and other giant herbivores were too large to have predators. Blaire Van Valkenburgh of the University of California, Los Angeles, and her colleagues analysed data on the relative body masses of modern predators and prey, and compared them with those of fossil specimens. They estimate that some Pleistocene predators, such as sabre-toothed cats and very large hyenas, were big enough to kill young megaherbivores — enabling them to control herbivore populations.

Proc. Natl Acad. Sci. USA **http://doi.org/8th (2015)**



NATURE COMMUN.

ACOUSTICS

Beads dance on sound waves

A bank of speakers can grip, move and rotate particles in air from one side (pictured).

Sound has been used to levitate small objects, but single-sided devices offered little manoeuvrability. Asier Marzo at the Public University of Navarre in Pamplona, Spain, and his colleagues used a flat array of 64 loudspeakers to levitate beads of polystyrene up to 3 millimetres wide. The authors used algorithms to create interference patterns in waves of ultrasound that formed regions of high and low intensity — shaped as tweezers, tornadoes or bottles — which allowed them to trap and then move the particles in various directions.

The device could be used to manipulate particles for targeted drug delivery or to operate tiny surgical devices from outside the body, say the authors.

Nature Commun. 6, **8661 (2015)**

ANIMAL BEHAVIOUR

Related wasps commit treason

Yellow-jacket wasps live to serve their mother, the queen, but will kill her if she fails to secure more than one mate.

KEVIN LOOPE

Colonies of yellow-jacket wasps (*Dolichovespula arenaria;* **pictured**) have a single queen that generates female workers, which rarely reproduce, and reproductive males. But Kevin Loope at Cornell University in Ithaca, New York, found that just under half of colonies eventually revolt, with the workers killing their queen and producing their own males. To find out why, Loope collected wasp nests and measured the workers' relatedness. Matricide was most common in colonies where workers were more closely related to each other.

This means that the queen had only one mate, making workers less closely related to the queen's sons than to the sons of other workers. Workers prefer males that are more closely related to them, so it benefits them to overthrow the queen and produce their own sons.

Curr. Biol. **http://doi.org/8vz (2015)**

CRYOSPHERE

Arctic open-water season grows

Ice could cover Arctic coastal regions for only half the year by the 2070s, if human-induced climate change continues.

Most of these areas are now covered in ice for more than half the year, and even all year in some places. Using data on daily sea-ice concentrations, Katherine Barnhart at the University of Colorado Boulder and her colleagues mapped changes in the Arctic's open-water season since pre-industrial times, and used models to project future changes. They found that throughout the Arctic, the season began to lengthen in the 1990s, with ice break-up starting earlier and freeze-up setting in later. In business-as-usual climate-change scenarios, the models indicate that the duration of open-water seasons for much of the region will start to exceed pre-industrial bounds by the middle of this century.

The expansion of the open-water season will affect all aspects of the Arctic environment that depend on sea-ice coverage, such as polar-bear foraging and the livelihoods of indigenous people, the authors say.
Nature Clim. Change **http://dx.doi.org/10.1038/nclimate2848 (2015)**

CANCER

Altered T cells hit pancreatic cancer

Genetically engineered immune cells that target a protein found on some pancreatic tumours can penetrate that cancer's defences, according to studies in mice.

Harnessing engineered T cells to combat cancer has been more successful for blood cancers than for solid tumours, such as those of the pancreas, which are protected by a dense cellular barrier and are particularly deadly. Philip Greenberg and Sunil Hingorani of the Fred Hutchinson Cancer Research Center in Seattle, Washington, and their colleagues engineered T cells to recognize a protein called mesothelin that is associated with the spread of certain pancreatic tumours. The engineered T cells were able to bind to this protein more tightly than did normal T cells.

The engineered cells infiltrated pancreatic tumours in mice, leading to an increase in tumour-cell death compared with control mice. Mice that received a series of engineered T-cell infusions lived nearly twice as long as those that did not.
Cancer Cell **http://dx.doi.org/10.1016/j.ccell.2015.09.022 (2015)**

IMMUNOLOGY

Worms conspire with gut microbes

Intestinal worms manipulate their host's immune system to ensure their survival, in part by changing the metabolism of the host's gut microbiome.

The worms, called helminths, infect around 2 billion people around the world, and are able to block harmful inflammatory responses in humans and mice. Nicola Harris at the Swiss Federal Institute of Technology in Lausanne and her colleagues studied mice infected with the helminth *Heligmosomoides polygyrus bakeri*, and found that mice that had been treated with antibiotics to kill gut bacteria before being exposed to the worms had more allergic airway inflammation than did untreated, worm-infected animals. Worm infection caused the microbiota to produce increased levels of short-chain fatty acids in mice, pigs and six out of eight human volunteers. The anti-inflammatory effects of worm infection were lost in mice that had been engineered to lack a receptor for the fatty acids.

The findings suggest that helminths and gut microbes have evolved this mechanism to regulate the host immune system over many millions of years, the authors say.
Immunity **http://dx.doi.org/10.1016/j.immuni.2015.09.012 (2015)**

DEVELOPMENTAL BIOLOGY

Survival boost for cloned embryos

Researchers have improved the success rate for producing cloned embryos or embryonic stem cells by removing a chemical group from DNA-binding proteins.

Transferring a nucleus from an individual's adult body cell into a human egg — a process called somatic cell nuclear transfer (SCNT) — could one day generate embryonic stem cells that match that person's DNA. But embryos made using SCNT rarely mature. To improve this, Dong Ryul Lee at CHA University in Seoul, Yi Zhang at Boston Children's Hospital in Massachusetts and their colleagues used a human messenger RNA encoding a protein that removes methyl groups from a type of histone protein found on DNA in the donor nucleus. When the authors injected the RNA into 56 human eggs that had received donor DNA, they found that 14.3% of the treated embryos developed into late-stage blastocysts, compared with none of the untreated controls.

Using this technique, the team derived embryonic stem cells from skin cells donated by people with age-related macular degeneration, which causes partial vision loss.
Cell Stem Cell **http://doi.org/8v2 (2015)**

SOCIAL SELECTION

Popular topics on social media

Funding of basic science stirs debate

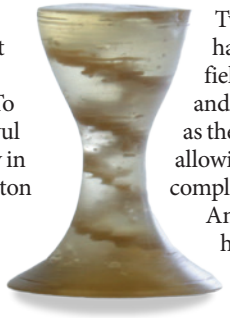
Pure science does not always stimulate innovation — rather, technological change often springs naturally from human inventiveness. Writer Matt Ridley makes this provocative point in a 23 October essay in *The Wall Street Journal* called 'The Myth of Basic Science' (go.nature.com/2bbqpg) that fuelled heated and thoughtful responses on social media about the role and benefits of science and technology. Ridley says that government-funded basic research is not the only path towards innovations that improve society.

But others countered that publicly funded research has many benefits. "The causes of technical and social change are manifold, and scientific research forms just part of the ecosystem, but this doesn't make it inconsequential," wrote Jack Stilgoe, a science-policy expert at University College London, in an article for *The Guardian* commenting on Ridley's essay (go.nature.com/zkkalt). Ridley responded to his critics on Twitter, saying that basic research is important but that government is not the only way to fund it.

NATURE.COM
For more on popular papers:
go.nature.com/tgdjzg

MATERIALS

Extra dimensions in 3D printing



Two research groups have used magnetic fields to tune the texture and strength of materials as they are being printed, allowing the formation of complex 3D structures.

André Studart and his colleagues at the Swiss Federal Institute of Technology in Zurich added magnetic particles at different concentrations to resins of varying viscosities. Applying a low magnetic field during the 3D printing process allowed the team to control the orientation of the particles, and hence the texture, within the printed object. The researchers used their technique to create a composite with an intricate internal spiral staircase (**pictured**). Their system could be used in robotics to print shape-changing objects that respond to environmental triggers, Studart says.

In a separate paper, Studart's former postdoc, Randall Erb, and his team at Northeastern University in Boston, Massachusetts, used the magnetic technique to improve the mechanical strength of 3D printed objects by controlling crack formation.
Nature Commun. **6, 8643 (2015); 8641 (2015)**

NATURE.COM
For the latest research published by *Nature* visit:
www.nature.com/latestresearch

NATURE COMMUN.

The expansion of the open-water season will affect all aspects of the Arctic environment that depend on sea-ice coverage, such as polar-bear foraging and the livelihoods of indigenous people, the authors say.
Nature Clim. Change **http://dx.doi.org/10.1038/nclimate2848 (2015)**

CANCER

Altered T cells hit pancreatic cancer

Genetically engineered immune cells that target a protein found on some pancreatic tumours can penetrate that cancer's defences, according to studies in mice.

Harnessing engineered T cells to combat cancer has been more successful for blood cancers than for solid tumours, such as those of the pancreas, which are protected by a dense cellular barrier and are particularly deadly. Philip Greenberg and Sunil Hingorani of the Fred Hutchinson Cancer Research Center in Seattle, Washington, and their colleagues engineered T cells to recognize a protein called mesothelin that is associated with the spread of certain pancreatic tumours. The engineered T cells were able to bind to this protein more tightly than did normal T cells.

The engineered cells infiltrated pancreatic tumours in mice, leading to an increase in tumour-cell death compared with control mice. Mice that received a series of engineered T-cell infusions lived nearly twice as long as those that did not.
Cancer Cell **http://dx.doi.org/10.1016/j.ccell.2015.09.022 (2015)**

IMMUNOLOGY

Worms conspire with gut microbes

Intestinal worms manipulate their host's immune system to ensure their survival, in part by changing the metabolism of the host's gut microbiome.

The worms, called helminths, infect around 2 billion people around the world, and are able to block harmful inflammatory responses in humans and mice. Nicola Harris at the Swiss Federal Institute of Technology in Lausanne and her colleagues studied mice infected with the helminth *Heligmosomoides polygyrus bakeri*, and found that mice that had been treated with antibiotics to kill gut bacteria before being exposed to the worms had more allergic airway inflammation than did untreated, worm-infected animals. Worm infection caused the microbiota to produce increased levels of short-chain fatty acids in mice, pigs and six out of eight human volunteers. The anti-inflammatory effects of worm infection were lost in mice that had been engineered to lack a receptor for the fatty acids.

The findings suggest that helminths and gut microbes have evolved this mechanism to regulate the host immune system over many millions of years, the authors say.
Immunity **http://dx.doi.org/10.1016/j.immuni.2015.09.012 (2015)**

DEVELOPMENTAL BIOLOGY

Survival boost for cloned embryos

Researchers have improved the success rate for producing cloned embryos or embryonic stem cells by removing a chemical group from DNA-binding proteins.

Transferring a nucleus from an individual's adult body cell into a human egg — a process called somatic cell nuclear transfer (SCNT) — could one day generate embryonic stem cells that match that person's DNA. But embryos made using SCNT rarely mature. To improve this, Dong Ryul Lee at CHA University in Seoul, Yi Zhang at Boston Children's Hospital in Massachusetts and their colleagues used a human messenger RNA encoding a protein that removes methyl groups from a type of histone protein found on DNA in the donor nucleus. When the authors injected the RNA into 56 human eggs that had received donor DNA, they found that 14.3% of the treated embryos developed into late-stage blastocysts, compared with none of the untreated controls.

Using this technique, the team derived embryonic stem cells from skin cells donated by people with age-related macular degeneration, which causes partial vision loss.
Cell Stem Cell **http://doi.org/8v2 (2015)**

SOCIAL SELECTION

Popular topics on social media

Funding of basic science stirs debate

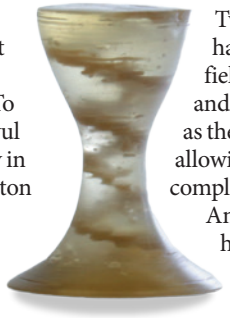
Pure science does not always stimulate innovation — rather, technological change often springs naturally from human inventiveness. Writer Matt Ridley makes this provocative point in a 23 October essay in *The Wall Street Journal* called 'The Myth of Basic Science' (go.nature.com/2bbqpg) that fuelled heated and thoughtful responses on social media about the role and benefits of science and technology. Ridley says that government-funded basic research is not the only path towards innovations that improve society.

But others countered that publicly funded research has many benefits. "The causes of technical and social change are manifold, and scientific research forms just part of the ecosystem, but this doesn't make it inconsequential," wrote Jack Stilgoe, a science-policy expert at University College London, in an article for *The Guardian* commenting on Ridley's essay (go.nature.com/zkkalt). Ridley responded to his critics on Twitter, saying that basic research is important but that government is not the only way to fund it.

NATURE.COM
For more on popular papers:
go.nature.com/tgdjzg

MATERIALS

Extra dimensions in 3D printing



Two research groups have used magnetic fields to tune the texture and strength of materials as they are being printed, allowing the formation of complex 3D structures.

André Studart and his colleagues at the Swiss Federal Institute of Technology in Zurich added magnetic particles at different concentrations to resins of varying viscosities. Applying a low magnetic field during the 3D printing process allowed the team to control the orientation of the particles, and hence the texture, within the printed object. The researchers used their technique to create a composite with an intricate internal spiral staircase (**pictured**). Their system could be used in robotics to print shape-changing objects that respond to environmental triggers, Studart says.

In a separate paper, Studart's former postdoc, Randall Erb, and his team at Northeastern University in Boston, Massachusetts, used the magnetic technique to improve the mechanical strength of 3D printed objects by controlling crack formation.
Nature Commun. **6, 8643 (2015); 8641 (2015)**

NATURE.COM
For the latest research published by *Nature* visit:
www.nature.com/latestresearch

NATURE COMMUN.

SEVEN DAYS *The news in brief*

RESEARCH

Reproducibility

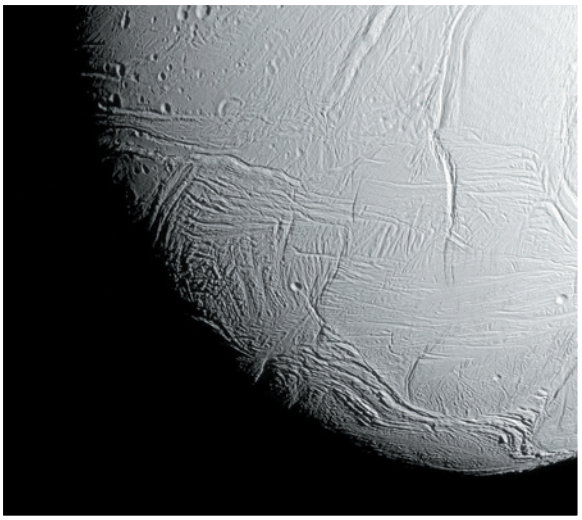
A suite of measures should be adopted to improve the reproducibility of biomedical research, according to a report released on 29 October by the London-based Academy of Medical Sciences. The report — produced with the backing of government funders and biomedical-research charity the Wellcome Trust — says that greater openness, preregistration of research protocols and better use of standards should all be considered, although there is no single cause of the problem of many studies being irreproducible. See go.nature.com/cwynyx for more.

Ozone-hole latest

This year's hole in the Antarctic ozone layer is the third largest ever observed, the World Meteorological Organization announced on 29 October. The hole's average size over 30 consecutive days spanning September and October was 26.9 million square kilometres, the largest on record after 2000 and 2006. The agency ascribes the increased size to colder-than-usual temperatures in the polar stratosphere. That drove the formation of more clouds on whose surfaces chlorine can readily convert to a form that destroys ozone. In the long term, the ozone layer is still expected to recover, because the 1987 Montreal Protocol phased out many chemicals that contribute to its destruction.

Chronic fatigue

The US National Institutes of Health is stepping up efforts to tackle chronic fatigue syndrome, also known as myalgic encephalomyelitis (CFS/ME). In an announcement on 29 October, the agency said that it would be centring its CFS/ME research programme in the National Institute of Neurological Disorders and Stroke. Its plans include a clinical study on its campus in Bethesda, Maryland, that will enrol patients with sudden-onset CFS/ME apparently caused by an infection.



NASA/JPL-CALTECH/SSI

Cassini dips into geysers

NASA's Cassini spacecraft took its deepest dive through the geysers spurting from Saturn's moon Enceladus on 28 October. The mission whizzed 50 kilometres above Enceladus's south pole (pictured, bottom), directly through the icy spray coming from an ocean of liquid water trapped beneath a thick layer of fractured ice. It was the most direct taste of the water that Cassini's chemical-analysis sensors will ever get; in the final fly-by in December, the spacecraft will bypass the geysers.

AWARDS

Maddox prize

The 2015 John Maddox Prize was awarded to Edzard Ernst and Susan Jebb on 3 November. Ernst, emeritus researcher at the University of Exeter, UK, was given the prize for his work on the truth, or lack thereof, in claims about complementary and alternative medicine. Jebb, a researcher at the University of Oxford, UK, received the prize for her work in furthering public understanding of nutrition. The prize for promoting science in the face of adversity is awarded jointly by *Nature* and the London-based charities the Kohn Foundation and Sense About Science. It is named after the late John Maddox, a former editor of *Nature*.

POLICY

One-child rule ends

All couples in China will in future be allowed to have two children, rather than one, the Communist Party announced on 29 October. But demographers predict little effect on population growth in China, where many women are more focused on a career than on having large families. The one-child rule was introduced in 1979 and is thought to have prevented almost half a billion births in a nation whose population now numbers 1.4 billion. In recent years the rule had been relaxed. See go.nature.com/skdr1n for more.

Pathogen rules

In the wake of a series of high-profile laboratory accidents in 2014, the White House issued a 187-page set of recommendations on 29 October for government agencies that work with dangerous pathogens. They include improvements to rules for reporting lab accidents and maintaining records.

Antarctic veto

The body that governs Antarctica's waters again failed to agree on plans for a protected area in the Ross Sea. The Commission for the Conservation of Marine Living Resources, meeting in Hobart, Australia, last week, has repeatedly considered the proposals but failed to reach the unanimous agreement among nations needed to create the area. The Antarctic Ocean Alliance, a coalition of non-governmental organizations, criticized the failure to protect the Ross Sea and another proposed area in East Antarctica.

GM opt-out block

The European Parliament has rejected a proposal that would allow European Union member states to restrict the importation of genetically modified (GM) feeds and foods that have been approved at EU level. In the 28 October vote, members argued that opting out of EU-wide agreements to allow the sale of GM food was incompatible with the EU's

single market. The European Commission tabled the proposal in April after it was agreed that EU member states could opt out of cultivating GM crops, which 19 of the 28 states have done so far.

FUNDING

Brain project

The European Commission signed a partnership agreement with the ambitious but controversial Human Brain Project (HBP) on 30 October. The agreement will take the project into its fully operational phase that begins next April, when the HBP will become an international organization intended to be a permanent infrastructure resource for neuroscientists. The management of the project has been modified following serious criticism by some neuroscientists during its start-up phase. See go.nature.com/qybrng for more.

Arecibo future

The US National Science Foundation (NSF) is seeking new management or new ownership for the Arecibo Observatory (**pictured**), it said in a 26 October notice. The future of the facility, the largest single-dish radio telescope on Earth, in Puerto Rico, has been in doubt for years. But the NSF, which provides roughly 75% of Arecibo's roughly

JEFF HITCHCOCK/CC BY 2.0

US$12-million budget, says that it is interested in options "that involve a substantially reduced funding commitment from NSF". Astronomers use the facility to study pulsars and the upper atmosphere and to help measure the risk posed by near-Earth asteroids.

POLITICS

Indian protest

Researchers in India have issued a warning over religious intolerance in the country. On 27 October, the Inter-Academy Panel on Ethics in Science, a body set up by the Indian National Science Academy in New Delhi, the Indian Academy of Sciences in Bangalore and the National Academy of Sciences in Allahabad, warned that recent events run counter to the country's constitutional requirement to "uphold reason and scientific temper". The statement follows the killing of three advocates of rational thinking, as well as other cases of violence linked to religious motives. An online petition voicing similar concerns was launched on 22 October. See page 20 for more.

PEOPLE

Digitized lives

Lauded bioinformatician Jun Wang, who stepped down in July from his post as chief executive of the world's largest genome-sequencing organization, BGI, in Shenzhen, has now launched his own company. Wang held an opening ceremony for the firm, called iCarbonX, in Shenzhen on 27 October. He says that the artificial-intelligence company will become a "Google for biotech" by collecting and analysing genomic, proteomic and other data from 1 million people. He plans to start recruiting within six months and to have a prototype platform in 3–5 years that will connect individual consumers, pharmaceutical companies, hospitals and other organizations.

Scientist sacked

Tsinghua University in Beijing confirmed to *Nature* on 2 November that it dismissed neuroscientist Zhang Shengjia following a controversy over a protein that senses magnetism. In September, Zhang reported manipulating neurons in worms by applying a magnetic field to the protein. A researcher at neighbouring Peking University who claims to have discovered the protein's magnetic-sensing capability and was in the middle of publishing his own results complained that Zhang had published his paper first. Tsinghua University has not yet specified a reason for Zhang's dismissal. Zhang denies that there is anything wrong with his paper, questions the procedure that led to his dismissal and says that he will file a rebuttal.

EVENTS

EPA versus VW

The US Environmental Protection Agency (EPA) has issued a second notice of violation against car manufacturer Volkswagen (VW) over allegations that the company installed a device to circumvent emission standards in some of its vehicles. The 2 November notice adds further car models to those listed on the notice from 18 September. VW previously admitted using 'defeat devices' to lower emissions during laboratory tests in some vehicles (see *Nature* http://doi.org/723; 2015).

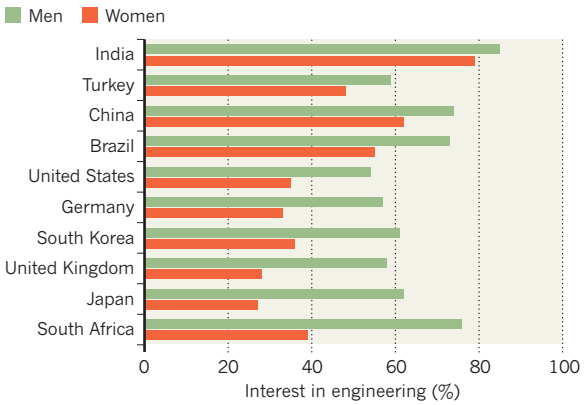
➲ NATURE.COM
For daily news updates see:
www.nature.com/news

TREND WATCH

Women in developing nations are challenging the gender bias often found in engineering. A survey in 10 countries, commissioned by the Queen Elizabeth Prize for Engineering Foundation, asked 10,000 people about their interest in engineering (see go.nature.com/khqpst). Overall, more men expressed interest than did women, but the gap was narrowest in emerging economies. In Britain, 28% of women and 58% of men showed interest, whereas the results for India were 79% for women and 85% for men.

GENDER BALANCE AND ENGINEERING

In developed nations, many more men than women are interested in engineering, in stark contrast to the situation in emerging economies.



SOURCE: QUEEN ELIZABETH PRIZE FOR ENGINEERING CREATE THE FUTURE (2015).

NEWS IN FOCUS

EROS HOAGLAND/REDUX/EYEVINE

Military-service members can suffer brain injury and memory loss when exposed to explosions in enclosed spaces, even if they do not sustain overt physical injury.

NEUROSCIENCE

Memory-enhancement trials move into humans

Research suggests that electrodes could compensate for damaged tissue.

BY SARA REARDON

A strategy designed to improve memory by delivering brain stimulation through implanted electrodes is undergoing trials in humans. The US military, which is funding the research, hopes that the approach might help many of the thousands of soldiers who have developed deficits to their long-term memory as a result of head trauma. At the Society for Neuroscience meeting in Chicago, Illinois, on 17–21 October, two teams funded by the Defense Advanced Research Projects Agency presented evidence that such implanted devices can improve a person's ability to retain memories.

By mimicking the electrical patterns that create and store memories, the researchers found that gaps caused by brain injury can be bridged. The findings raise hopes that a 'neuroprosthetic' that automatically enhances flagging memory could aid not only brain-injured soldiers, but also people who have had strokes — or even those who have lost some power of recall through normal ageing.

Because of the risks associated with surgically placing devices in the brain, both groups are studying people with epilepsy who already have implanted electrodes. The researchers can use these electrodes both to record brain activity and to stimulate specific groups of neurons. Although the ultimate goal is to treat traumatic brain injury, these people might benefit as well, says biological engineer Theodore Berger at the University of Southern California (USC) in Los Angeles. That is because repeated seizures can destroy the brain tissue needed for long-term-memory formation.

Short-term memories are thought to be created when a part of the brain called the hippocampus aggregates sensory information, as well as the perception of space and time, and holds it readily accessible for a short while. Accessing the memory during that time will solidify it into a long-term memory. ▶

Key to this process is a signal that travels from one part of the hippocampus called CA3 to another, called CA1. Berger and his colleagues hypothesize that recreating that signal might restore the ability to solidify memories in people with damage to the hippocampus.

In one of the studies presented at the Chicago meeting, researchers asked 12 people with epilepsy to look at pictures and then recall up to 90 seconds later which ones they had seen. While the participants did this, the researchers recorded the firing patterns in both CA3 and CA1.

They then developed an algorithm that could use the activity of the CA1 cells to predict the pattern that was coming from CA3. Compared with the actual patterns, their predictions were accurate about 80% of the time.

By using this algorithm, the researchers should be able to stimulate the CA1 cells with a pattern that mimics an appropriate CA3 signal even if a person's CA3 cells are damaged, Berger says. In previous studies on monkeys trained to do the picture-recall task, receiving a juice reward when correct, his group has shown that stimulating CA1 with an appropriate pattern significantly improved the animals' performance (R. E. Hampson *et al. J. Neural Eng.* **10,** 066013; 2013).

USC biomedical engineer Dong Song, a member of the team, says that the group has tried the stimulation on a woman with epilepsy, but that it is too early to know whether it has improved her memory. He says that the researchers plan to apply it to more people in the coming months. Eventually, a device might be developed that would detect when the hippocampus is not efficiently encoding short-term into long-term memory and provide stimulation to support the process.

It is amazing that the memory-formation code can be so accurately predicted, says neurobiologist Howard Eichenbaum at Boston University in Massachusetts. But he cautions that mimicking it could be difficult if the CA1 cells are so badly damaged that they will not respond properly to stimulation. And he adds that because the hippocampus is so complex and receives inputs from many connections in the brain, stimulating it with the CA3 signal alone may not be enough.

It is amazing that the memory-formation code can be so accurately predicted.

Thomas McHugh, a neuroscientist at the RIKEN Brain Science Institute in Tokyo, says that he has been following the team's work for years and has been consistently surprised at how well the approach has worked in animal models. "The data is convincing, but I'm still at a loss for understanding," he says. Many parts of the brain are organized in obvious ways: in the motor cortex, for example, stimulating a particular spot causes motion in a specific part of the body. But there is no such obvious organization in the hippocampus, so it is unclear why stimulating certain locations leads to predictable results.

A team at the University of Pennsylvania (Penn) in Philadelphia is taking a different approach to enhancing memory that requires an even less detailed understanding of how the process works.

The team exploits the fact that people's memory skills fluctuate over time depending on variables such as how much caffeine they have consumed or whether they are under stress. The team has found, again by working with people with epilepsy, that stimulating a region called the medial temporal lobe, which houses the hippocampus, improves memory that is functioning poorly. But when memory is functioning well, stimulation impedes it.

In a study that they presented at the Chicago meeting, Penn neuroscientist Daniel Rizzuto and his colleagues recorded brain activity in 28 people as they recalled a list of words. Using these patterns, the researchers developed an algorithm that predicted with high accuracy whether a person would remember a given word. By stimulating the brain only when a person read words that were likely to be forgotten, the researchers could boost performance by up to 140%.

Penn psychologist Michael Kahana says that the team has recorded from the brains of about 80 people in total and is seeking regulatory approval to use a more precise electrode array.

Although it would be useful from a basic-science viewpoint to discover why stimulation works so well, McHugh says, it may be worth developing therapies based on it even if it is not fully understood — as long as it can be proved to be safe and effective. ■

SPACE

Historic Rosetta mission to end with crash into comet

There were other options, but super close-up shots on descent will provide science bonanza.

BY ELIZABETH GIBNEY

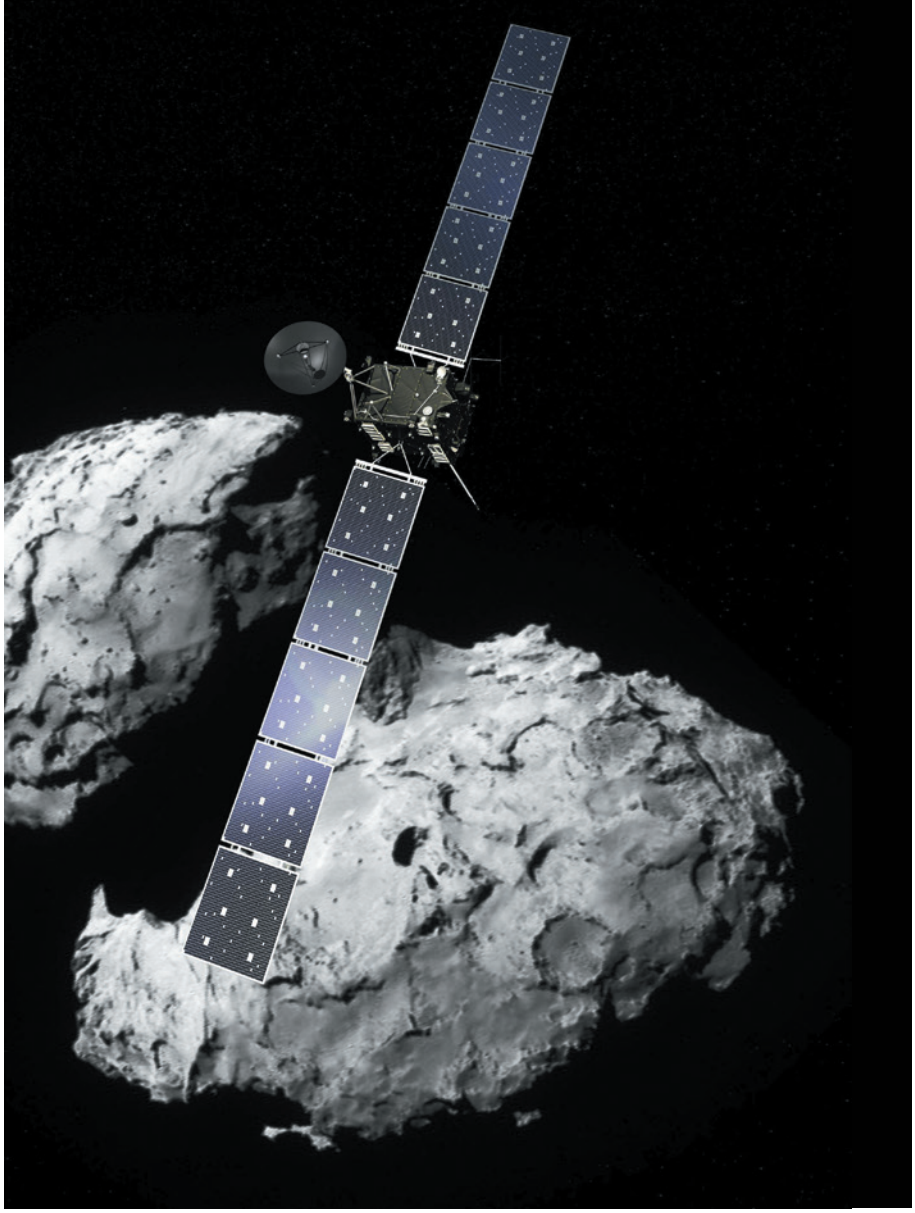
A year since a probe called Philae made history by touching down on a comet, the team that pulled off the feat is plotting a different kind of landing. Next September, the European Space Agency will crash Philae's mothership Rosetta into the icy dust ball, but as gently as possible.

The dramatic act will bring the mission to an abrupt end — and give Rosetta's wealth of sensors and instruments their closest view of the comet yet. "The crash landing gives us the best scientific end-of-mission that we can hope for," says Rosetta project scientist Matt Taylor.

The collision will be emotional for the scientists, some of whom have worked on the mission since its inception in 1993. "There will be a lot of tears," says Taylor.

Launched in 2004, the Rosetta orbiter caught up with the comet 67P/Churyumov-Gerasimenko ten years later as the rock was travelling from deep in space towards the Sun — and dropped Philae onto the surface a few months later, on 12 November. Scientists have not heard from Philae since July, and don't know if they will do so again, but Rosetta's operations to survey the comet from orbit are in full swing. However, the orbiter can't keep up this work indefinitely. Funding for the mission runs out in September 2016 — and by that time 67P/Churyumov-Gerasimenko will be well on its way back out into deep space, where the solar-powered orbiter will receive too little sunlight to function.

Discussions about what to do with Rosetta when that happens have continued for more than a year. Rosetta flight director Andrea Accomazzo says that, ideally, Rosetta would hibernate while the comet remains in deep space, then be resurrected when 67P again approaches the Sun in 4 or 5 years' time. But the cold of deep space would probably damage the craft, Accomazzo says; others fear that fuel and other resources would run out. Moreover,



SPACECRAFT: ESA/ATG MEDIALAB; COMET: ESA/ROSETTA/NAVCAM

Artist's impression of the Rosetta orbiter approaching the comet 67P/Churyumov-Gerasimenko.

many of the mission's principal investigators (PIs) began their work more than 20 years ago and "there's no point putting an old experiment with old PIs into hibernation", jokes Kathrin Altwegg, a planetary scientist at the University of Bern.

Crash-landing Rosetta emerged as the preferred option last year, but only now are orbiter navigators and operators working out how to go about it. Rosetta's closest encounter with the comet so far was from 8 kilometres above the surface, when it dispatched Philae. The current thinking sees Rosetta spiral down to a similar distance next August before creeping ever closer in elliptical orbits and crashing in September, says mission manager Patrick Martin — but that could still change.

Although Philae sent back some data during its descent, Rosetta has more powerful — and more varied — sensors and instruments. The orbiter will also descend much more slowly than Philae did, allowing it to gather more data and better pictures. Once it gets to 4 kilometres, for example, Rosetta should be able to distinguish between the gases emerging from each of the duck-shaped comet's two lobes to determine whether the regions vary in composition, says Altwegg, who leads the team behind ROSINA (the Rosetta Orbiter Spectrometer for Ion and Neutral Analysis). That could shed light on the environments in which each was formed.

Rosetta's cameras will get their best-resolution shots of the comet's surface yet — less than 1 centimetre per pixel once the craft is within 500 metres of the surface, adds Holger Sierks, PI for Rosetta's OSIRIS (Optical, Spectroscopic, and Infrared Remote Imaging System). This will allow researchers to look at surface properties and link these to comet activity that Rosetta has observed from orbit.

OVER AND OUT

How far into the descent Rosetta will be able to send data back to mission control will depend on whether engineers can design the final trajectory such that the craft crashes on the side of the comet that faces Earth. Navigating while close to the comet will be difficult because the body's gravitational field is uneven, but spacecraft-operations manager Sylvain Lodiot hopes that the orbiter will transmit until the very end.

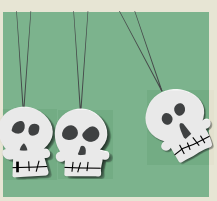
The crash will definitely be a hard stop to the mission, he says, however gentle the landing. Designed to manoeuvre in orbit, once Rosetta is on the comet's surface it will no longer be able to point its antenna to communicate with Earth. Similarly, it will not be able to angle its solar array, so it will lose power, says Lodiot. "Once we touch, hit or crash, whatever you want to call it, it's game over."

Before then, though, the mission still has much to accomplish. As the comet approached the Sun, it heated up, with vaporizing ice causing more and more gas and dust to stream from its surface. Rosetta had to retreat into a wider orbit to stop the dust from confusing its navigation system. But now that the comet is speeding away from the Sun, mission scientists are relishing the opportunity to steer Rosetta back in. Priorities will then be to get images that would enable comparisons of the comet before and after its swing around the Sun, as well as a close-up of the southern hemisphere, which was largely in darkness until May and will disappear back out of view in March.

Rosetta will also resume listening out for Philae. Given the huge public interest in anything to do with the lander, Rosetta's finale will make for a fitting end to the story, adds Altwegg. "This way Rosetta gets to live happily ever after on the comet with Philae." ■

MORE ONLINE

TOP STORY



Zombie physics: six baffling results that just won't die go.nature.com/5kfcib

MORE NEWS

- Grant application rejected over choice of font go.nature.com/g2h1fb
- Artificial-intelligence institute launches free science search engine go.nature.com/tgilfn
- Troubled brain project gets another 3 years' funding go.nature.com/qybrng

Q&A

ROBERT TAYLOR

How to make biomedical research more reproducible go.nature.com/cwynyx

BAY ISMOYO/AFP/GETTY

Raging forest fires threaten orangutans such as this one at a rehabilitation centre in Borneo.

CONSERVATION

Scramble to save Borneo's orangutans

Fuelled by El Niño and land-management blunders, Indonesian fires are consuming precious habitat.

BY NADIA DRAKE

The world's only wild orangutans — already besieged by logging, hunting, pet trading and the steady expansion of palm-oil plantations — are now threatened by forest fires that have burned for months on the islands of Borneo and Sumatra in southeast Asia. In the toxic smoke and haze, locals and researchers are scrambling to protect the estimated 50,000 remaining orangutans that live only on those two islands.

Fires erupt every year in Indonesia during the dry season, as farmers, plantation owners and others deliberately burn forest to clear land or to settle territorial disputes. But this year's El Niño weather pattern, combined with a legacy of land-management practices that have dried the soil and degraded vast swathes of peat-swamp forest, turned this burning season into an environmental catastrophe that has destroyed more than 2 million hectares of forest throughout Indonesia, to which Sumatra and much of Borneo belong.

Since late summer, teams of researchers have headed out from the city of Palangkaraya in Borneo to find and fight new blazes. Some patrol the rivers and others head into the forest, where extinguishing the flames can require drilling more than 20 metres down to reach the water table — tough, gruelling work that is carried out amid tropical heat and in a persistent, menacing orange haze.

One day in October, Simon Husson, director of the UK-based Orangutan Tropical Peatland Project, deployed a drone at the Borneo Orangutan Survival Foundation's centre for orangutan rescue and rehabilitation near Palangkaraya. "Eyes in the sky are a huge help," he says. "On the ground, you're in choking smoke and the haze is severely restricting visibility."

As the drone rose above the smoggy blanket, its camera glimpsed a new fire burning deep in the forest. The fire was remote enough not to threaten the orphaned and injured orangutans being readied for reintroduction to the forest, "but you can't help thinking about the wild ones out there", Husson says.

Husson and his colleagues have temporarily abandoned their normal research activities in the 6,000-square-kilometre Sabangau Forest, which is home not just to orangutans but also to rare Bornean white-bearded gibbons, sun bears and pangolins, to help local fire-fighting teams with cash and personnel. "Not only is [research] pretty unimportant right now," he says, "it's basically impossible to study the orangutans in the canopy as we can't see them for the smoke."

Peat fires devastate orangutan populations primarily by destroying crucial habitat, but the animals are also susceptible to the same types of smoke- and haze-induced respiratory problems as humans. The charismatic arboreal apes are already endangered throughout their range; their population is estimated to have declined by 78% from more than 230,000 a century ago. "Over half the world's orangutans live in peat-swamp forests, and every one of these peatlands in Borneo right now is on fire, somewhere," Husson says.

Undisturbed peat forests are actually incredibly fire resistant, says Susan Page, a geographer at the University of Leicester, UK, who studies peatlands in southeast Asia, because the swamps are damp enough to make ignition difficult. But, unfortunately, large tracts of Borneo's peatland are anything but undisturbed. In 1996, Indonesia's then-president Suharto launched the Mega Rice Project, which tried to transform 1 million hectares of Bornean peatland into rice paddies. Draining the peat was essential for the plan, and despite the fact that no rice was ever harvested, canals that were cut through the forests have been draining water from the peat ever since.

The infernos in Indonesia have climate implications as well. Normally, Borneo's peat forests are efficient carbon stores, holding tonnes of organic matter in layers of compressed plant material that can be more than 15 metres thick. But when that peat burns, the accumulated carbon is released. This year, the fires have already released more than 1.5 billion tonnes of carbon dioxide into the atmosphere — more than Japan's annual carbon emissions. Since September, carbon emissions due to the fires have exceeded the daily production of the United States on at least 38 days, prompting one conservation scientist to call this year's fires the "biggest environmental crime of the twenty-first century".

The situation is unlikely to get better without an extended period of rain or a serious commitment from the Indonesian government. If the El Niño-driven drought persists, as some climate models predict, this year's fire season could last well into 2016.

"Severe fires did not occur before there was intensive land-use development," Page says. "Solutions will require strong political leadership and investment." ■

BUSINESS

Tech investors bet on synthetic biology

Once hesitant, Silicon Valley venture capitalists are warming to the idea of engineered cells.

BY ERIKA CHECK HAYDEN

MONEY FOR MICROBES
Investments in synthetic-biology start-ups have increased dramatically in the past three years. Much of the funding comes from prominent technology investors.

COMPANY	YEAR FOUNDED	BUSINESS	TOTAL FUNDS (US$)	NOTABLE INVESTORS
Twist Bioscience	2013	DNA synthesis	$82.11 million	Yuri Milner (Internet-company investor)
Zymergen	2013	Microbial-strain optimization	$44 million	Obvious Ventures; Eric Schmidt (Alphabet executive chairman)
Ginkgo Bioworks	2008	Microbial engineering	$54.12 million	Matt Ocko (Facebook and Zynga investor)
Bolt Threads	2009	High-performance fibres	$40 million	Peter Thiel and Max Levchin (PayPal co-founders)
Transcriptic	2012	Robotics for biology labs	$14.37 million	Jerry Yang (Yahoo co-founder)
Riffyn	2014	Software	$1.8 million	O'Reilly AlphaTech Ventures
Emerald Therapeutics	2010	Technology platforms	$34 million	Peter Thiel and Max Levchin

SOURCE: CRUNCHBASE.COM

In 2012, Emily Leproust was trying to raise money to start Twist Bioscience, a company that aimed to synthesize DNA more quickly and more cheaply than existing methods allowed. But many investors were spooked by the perception that synthetic biology — the engineering of microorganisms to make useful products such as drugs, food ingredients and materials — would not turn a profit. "It was a lonely time," Leproust recalls.

Three years later, Silicon Valley's big fish — technology investors with billions of dollars at their disposal — have finally ventured into synthetic biology's small pond. Scared away from conventional biotechnology in past by the risky and expensive prospect of drug development, they are now lured by what they see as synthetic biology's huge market potential, plummeting operating costs, improved business models and an increasing emphasis on computing.

"The toolkit is evolved from where it was two years ago; synthetic biology is going through a digitization and automation," says Nan Li, principal at investment firm Obvious Ventures in San Francisco, California. Li has coined the term 'biobiotics' for the current state of synthetic biology. "We see that this looks a lot more like a data and software problem, and we can understand that and get excited about that."

Software tools and robotics have reduced the cost of all parts of the process, from creating a genetic 'program' to inserting it into a microbe and testing it in the lab. For instance, synthetic-biology start-up firm Zymergen in Emeryville, California, uses machine learning — computer algorithms that evolve in response to data — to guide the engineering of fungi and bacteria that perform industrial processes more efficiently. The company also depends heavily on robotic automation of its labs, reducing the need to pay human workers.

"Synthetic biology is going through a digitization and automation."

Zymergen and Twist Bioscience are among a global class of synthetic-biology firms that have raised a record-breaking US$560.7 million this year, including $227 million for companies that use the wildly popular CRISPR/Cas9 gene-editing technology, says John Cumbers, founder of SynBioBeta industry group (see 'Money for microbes'). Overall, 24 newly created synthetic-biology companies have raised funding in 2015, compared to fewer than 6 in 2012.

Much of the work that goes into developing a synthetic-biology product is shifting to computers and robots. This means that as the sector grows, every dollar invested can produce more progress. "That reduces the cost for the venture capitalist who is having to put up the money," says Matt Ocko, co-managing partner of the Data Collective fund and one of several investors who will speak at the SynBioBeta industry conference in San Francisco on 4–6 November.

PAST MISTAKES

The new generation of synthetic-biology companies also has the advantage of experience. Many of today's founders are alumni of firms that struggled because they set their sights on huge, highly regulated industries that were hard to break into — such as pharmaceuticals or fuel.

Now, by contrast, start-up companies are focusing on niche areas where they can quickly bring products to market, such as speciality chemical, food, cosmetics and clothing industries — while hoping that opportunities will emerge to tackle other, bigger targets.

"People have come up with much more clever ways of generating revenue much faster," says Derek Greenfield, a co-founder of Industrial Microbes, which aims to engineer yeast that can synthesize chemicals using methane as a raw material. Bolt Threads in Emeryville, for example, is making fabrics from yeast by engineering the cells' metabolic pathways to mimic the processes used by spiders to make silk.

Other companies aim to use microorganisms to make rubber, egg proteins, rhino horn, vanilla flavouring, rose-scented extract or coffee more cheaply, more ethically or of higher quality. For example, coffee fermented by engineered microbes could replace a high-end brew made with beans harvested from civet faeces under conditions that some consider inhumane.

Long term, synthetic-biology businesses have the potential to achieve something more meaningful — and profitable — than another social-media site or cloud-storage service, says Jason Kelly, co-founder of microbial-engineering company Ginkgo Bioworks in Boston, Massachusetts. Like electric-car maker Tesla or commercial spacecraft company SpaceX, he says, synthetic biology could revolutionize economic sectors that are ripe for innovation — or even create new industries.

Li concurs. "As we understand more about the biology and limitations of these microbes, there's a potential to create entire new product categories," he says. "A lot of limitations that we take for granted can be stretched or pushed; there's just a lot more levers to pull." ■

INDIA

Scientists decry killings of secularists

Indian academy members condemn intolerance.

BY T. V. PADMA

Indian scientists are voicing concerns over religious intolerance and the killings of three noted advocates of rational thinking. The actions are unusual in a country where scientists rarely comment on political issues, says physicist Shri Krishna Joshi, a member of India's Inter-Academy Panel on Ethics in Science.

Anti-superstition activist Narendra Dabholkar was killed in 2013, communist politician Govind Pansare in February this year and literature scholar Malleshappa Kalburgi in August. All three deaths have been blamed on members of extreme right-wing Hindu groups.

On 22 October, scientists launched an online petition to India's president, Pranab Mukherjee, protesting against the killings. "The government has failed to check or discourage the anti-rational environment," says petition signatory Naresh Dadhich, a physicist at the Inter-University Centre for Astronomy and Astrophysics in Pune, India.

The petition was followed on 27 October by a statement from the Inter-Academy Panel on Ethics in Science, set up by the Indian National Science Academy in New Delhi; the Indian Academy of Sciences, Bangalore; and the National Academy of Sciences in Allahabad. The Indian constitution mandates that "its citizens abide by and uphold reason and scientific temper", the statement said. Several statements and actions "run counter to this constitutional requirement," it notes.

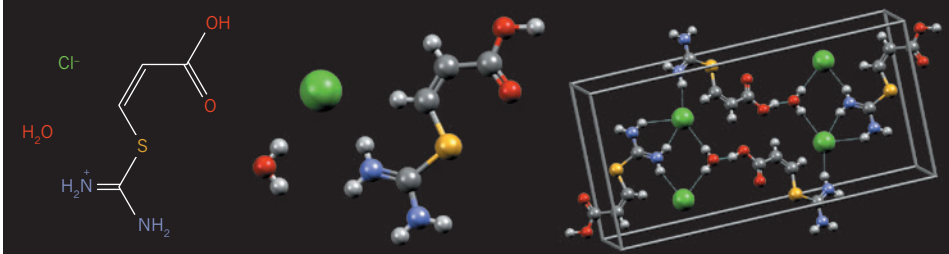
Indira Nath, a member of the panel and an immunologist at the Indian National Science Academy, says that the panel wants to "bring back rationality and scientific thinking to the mainstream".

Also last week, more than 100 scientists from leading Indian institutes, including national award winners, three fellows of the Royal Society in London, and a foreign associate of the US National Academy of Sciences, signed a second statement expressing deep concern over the "climate of intolerance".

Pushpa Mittra Bhargava, former director of the Centre for Cellular and Molecular Biology in Hyderabad, says that he plans to return a national award in protest. "Science is about reason and rationality. If three rationalists can be killed, scientists too can be killed." ■

CRYSTAL CHALLENGE

The 3D structure that a molecule adopts in a crystal is very difficult to predict — but defines what properties the molecule has.



The structural formula of a molecule reveals which atoms are connected at a 2D level.

Chemists are making progress at predicting how complex molecules will assemble in 3D space — there are millions of possibilities.

The 3D orientation repeats in a crystalline lattice with a structure that dictates the molecule's mechanical, chemical and physical properties.

CHEMISTRY

Software predicts crystal structures

Chemists have succeeded at a fiendish task — forecasting how complex molecules will assemble in 3D.

BY ELIZABETH GIBNEY

Sketch the structure of an organic molecule on a napkin and it may not be apparent that there are millions of possible ways that it could assemble as a 3D crystal. Now, a collaboration of dozens of chemists and computer programmers has successfully predicted the crystal structure of five, complex, 'drug-like' organic molecules — using nothing but a 2D map showing which atoms connect to which.

The achievement, announced on 27 October at a workshop in Cambridge, UK, paves the way for software that would cut the cost of the design and manufacture of drugs and other chemical products, as well as further our understanding of fundamental chemistry.

A molecule's crystal structure determines its properties (see 'Crystal challenge'). In 1998, the US pharmaceutical firm Abbott Laboratories learned this the hard way when it had to pull production of the capsule form of the HIV treatment ritonavir because the drug had started switching to an unexpected structure during manufacture. The crystal structure that a molecule adopts is generally the one with the lowest energy, but predicting what this is for any particular molecule is "fiendishly difficult", says Colin Groom, executive director of the Cambridge Crystallographic Data Centre (CCDC).

Even when chemists know which atoms are connected to which, the atoms can still be in different orientations because the bonds that connect them can bend and rotate in myriad ways. There are also multiple options for how molecules can pack together. "It is like looking for a needle in an unimaginably big haystack," says Anthony Reilly, a structural chemist at the CCDC.

Since 1999, the CCDC has organized six challenges known as the Blind Test of Organic Crystal Structure Prediction Methods. Rather than a contest, organizers see the challenge as a large collaborative attempt to compare the strengths of the latest techniques. "The groups participating represent pretty much the entire crystal-structure prediction community, and the methods used are the very best developed," says Groom.

The challenge typically takes place over a year, and sets two major problems. First, teams must come up with a list of all possible arrangements in which the molecules could form a crystal. Some teams do a rough calculation of the energy of each to whittle down the list, burning up hundreds of thousands of hours of computing time; others start with pure guesses and iteratively 'breed' the most stable to derive possible candidates more quickly. In the second stage, teams take the shortlists — sometimes assembled by a different group — and do more-precise

calculations of the energy of each, producing a ranking of the candidates.

The latest challenge, which included a record 25 teams — ten more than the previous contest in 2010 — brought a "massive improvement", says Groom. The molecules selected were "nasty, real-life systems" of the size and complexity to be interesting to drug companies. Previous challenges had included molecules that were flexible or made from multiple parts. This year's challenge combined such features in the same molecules and for one target, asked participants to predict not just one stable structure, but all its many stable forms, known as polymorphs.

PROBLEM SOLVED

The teams rose to the challenge: at the Cambridge workshop, the CCDC announced that each of the five targets, and their polymorphs, appeared in at least one of the shortlists produced by the various methods. A paper with the full results will be published in a special issue of *Acta Crystallographica Section B*.

Moreover, one team, led by Marcus Neumann at the German company Avantgarde Materials Simulation in Freiburg, included the correct solution in each of its shortlists. Had the team combined its efforts with those of a group — led by theoretical chemist Alexandre Tkatchenko at the Fritz Haber Institute in Berlin — that got a perfect score in the ranking phase, the two would together have achieved a perfect score for both rounds and across all targets. Such a result has never occurred in the history of the contest. "With what you have seen from me, and what you have seen from Tkatchenko," says Neumann, "it is fair to claim that to a large extent, this blind test has shown that the problem of organic crystal structure prediction has been solved."

"We have finally kicked the user out of the equation."

More so than in previous blind tests, teams including both Neumann's and Tkatchenko's took into account how quantum mechanical interactions would contribute to the energy of structures. In particular, Tkatchenko used a method published just last year that encompassed these interactions over longer ranges than has been done previously. And Neumann says that his program was unique because it made every decision by itself; most others required human decisions once the computer had returned its calculations. "We have finally kicked the user out of the equation," Neumann says.

Although others agree that the joint feat is a milestone, they stop short of declaring the problem of crystal structure prediction solved. "This does not mean that they would have cracked the problem of predicting all organic crystal structures," says Sally Price, a theoretical chemist at University College London.

And some are frustrated that Neumann has refused to release his computer code: "The day I have a pension plan, I will talk about this freely," he told the workshop. That will make it hard for others to build on his team's breakthrough. "We don't really have a sense of how it works," says challenge participant Claire Adjiman, a chemical engineer at Imperial College London. "But I understand why he doesn't tell us more."

Tkatchenko and Neumann now plan to work together. "My own interest is to understand polymorphism and be able to offer tools to people," says Tkatchenko. "His interest is more commercial, but I'm sure we can find the middle ground."

Both Price and Neumann, meanwhile, are already working with industry on how to use their prediction calculations in drug development. ■

CORRECTION

The News story 'Vaccine gets cautious boost' (*Nature* **526,** 617–618; 2015) incorrectly stated that David Kaslow was involved in the early development of RTS,S.

THE BABY EXPERIMENT

BY LINDA GEDDES

WES FERNANDES/*NATURE*

A London lab is deploying every technology it can to understand infant brains, and what happens when development goes awry.

Baby Ezra is sitting on his mother's lap and staring at the computer screen with the amazement of someone still new to the world. The five-month-old's eyes rest on a series of pictures: three dancing women, four black circles, then a face among random objects. Ezra studies the screen with fascination — although now and then, his attention wanders. He lets out a gurgle, and moments later, a short cry. He is chewing a sock.

Below the screen, a box is shining infrared light at his cornea, and then capturing and processing the reflected light to work out the direction of his gaze. Behind a curtain, postdoc Jannath Begum Ali checks the data streaming in on her monitor. This set-up is part of a sophisticated experiment to understand the early development of the human mind in the Babylab at Birkbeck, University of London. The scientists here will closely monitor Ezra's brain and behaviour at visits over the next two and a half years.

Oblivious to his important role in science, Ezra furrows his brow into a frown. What happens next is apparent only to his mother, who turns him around and checks his behind. With just half of a planned 15-minute observation complete, Ezra has defecated. At that point, everyone takes a break.

At Babylab, a 6-month-old has her brain's electrical activity monitored.

How do you get into the mind of a human being who cannot speak, does not follow instructions and rudely interrupts your experiments? That is the challenge embraced by scientists at the Babylab. The brain undergoes more change during the first two years of life than at any other time: consciousness, traits of personality, temperament and ability all become apparent, as do the first signs that development could be drifting off course. But this period is also the most difficult to explore, because many of the standard tools of human neuroscience are useless: babies will not lie awake and still in an imaging machine, and they cannot answer questions or do as they are told. Researchers have measured infants' interest and attention mostly by tracking their gaze — but even this method has been criticized as crude.

"There are many studies where someone tries to prove that the baby understands goals, causality, number — and in 99% of those studies the only measure they look at is a change in looking time," says Jerome Kagan, a psychologist at Harvard University in Cambridge, Massachusetts.

The field is now becoming more sophisticated, thanks in part to the Birkbeck lab. Scientists there have pioneered techniques such as infant near-infrared spectrometry (NIRS), which measures brain activity by recording the colour, and therefore the oxygenation, of blood. They are also trying to strengthen conclusions by combining multiple techniques. Among the handful of baby labs around the world, this makes the London one stand out. "They are doing research on babies using every single technique you could imagine," says Richard Aslin, an infant-behaviour researcher and director of the Rochester Center for Brain Imaging in New York.

> "They are doing research on babies using every single technique you could imagine."

The lab has used such tools to reveal a series of 'firsts' about the infant mind: that babies prefer to look at faces that are looking directly at them, rather than away from them; that they respond to such direct gaze with enhanced neural processing[1]; and that changes in this brain response may be associated with the later emergence of autism — the first evidence that a measure of brain function might be used to predict the condition[2]. In 2013, the Babylab started the flagship project of which Ezra is part: an effort to study infants from 12 weeks old who are at high risk of autism spectrum disorder or attention deficit hyperactivity disorder (ADHD), alongside a control group, in order to detect more early signs of these conditions and find behavioural therapies that might help. "It's an exciting, and emerging, field," says Mark Johnson, director of the Babylab.

And, like its subjects, the London lab is growing up. In 2014, Johnson received £2.3 million (US$3.5 million) from a trio of foundations to establish a toddler lab at Birkbeck, in which children aged 18 months to 3 or 4 years old will be attached to wireless forms of electroencephalography (EEG), NIRS and eye-tracking technology as they walk around, play and interact with other children. The aim is to understand the brain during toddlerhood, the time when children start to appreciate the difference between self and other, complex language develops and long-term memories are first laid down. "In child development in general, but also in our brain-development work, the terrible twos are a major black hole," Johnson says.

LOOK AND LEARN

There is a well-worn adage in show business that you should never work with children or animals. Johnson built his career doing both. For his PhD project in the 1980s, he investigated whether day-old chicks formed social attachments to any object placed in their pen, or if they preferred ones that resembled a mother hen. (The chicks were particularly drawn to objects with hen-like necks and faces, but weren't too fussy about the rest of their looks[3].) But Johnson was more interested in human development, so after his PhD he took a research-scientist position in London to begin studying infants. "In some ways that's not as big a jump as it sounds," he says. "In both cases you're trying to develop tasks and get information from non-verbal creatures."

Scientists have been attempting practical research with babies since the middle of the twentieth century. One of the first to do so was Jean Piaget, a Swiss psychologist who used detailed observations of infants and older children to gain insight into how they understand the world — including, famously, by hiding an object to see whether infants try to find it. He concluded that babies cannot grasp the concept that an object still exists when it is out of sight until they are around eight months old. Piaget went on to develop the theory that babies are essentially born as blank slates, but possess the machinery that motivates them to explore the world and allows them to assimilate knowledge.

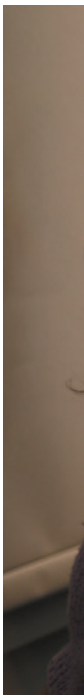
Infant neuroscience leapt forward in the early 1960s, when the US developmental psychologist Robert Fantz started measuring the amount of time babies spent looking at something as a way to gauge how interested in it they were. Fantz reported that a two-month-old baby spent twice as long looking at a sketch of the human face as at a bullseye, for instance. Experiments based on gaze measurements have been the field's workhorse ever since. "It is no exaggeration to say that without looking-time measures, we would know very little about nearly any aspect of infant development," says Aslin. Gaze experiments have led some researchers to conclude that, far from being blank slates, babies are born with an innate appreciation of number and human faces, as well as the ability to recognize when their mother's native language is being spoken — a familiarity proposed to develop through hearing speech while in the womb.

"There have been literally thousands of experiments done with these looking-time methods," Aslin says, "and by and large it is a pretty reliable technique; you can have two labs running the same experiment and you get the same results." But Aslin and Kagan are two of a growing number of researchers who think that such infant studies should be viewed with caution: it can be dangerous to infer too much about the workings of a baby's mind from just their fleeting glance — and they worry that some labs do not control for confounding factors as well as they should. "Looking time is under the control of so many conditions," Kagan says. "What are the physical features of the stimulus? Are its lines mainly curved or straight? What colours are present? How much contrast in lighting is there?"

Babies' brains are growing and developing at an extraordinary pace, which makes comparisons between different ages difficult: a newborn's gaze might reflect innate abilities, but a seven-month-old's will also be influenced by what he or she is starting to learn and remember about the world. "An infant may look longer in order to relate the event to what it already knows," says Kagan. "The main point is that no single measure is able to supply all the evidence required for conclusions about what infants know."

That was the opinion that Johnson quickly reached when he began infant research: the reliance on looking time and observations alone were unsatisfying. He established a baby lab at University College London (UCL) in 1993, and it moved to more spacious premises at Birkbeck in 1998. From the start, Johnson wanted to take a more high-tech approach to investigating brain development than were the handful of other similar labs.

In 2005, Johnson and his colleagues combined observations of looking time with electrical measurements of brain activity to investigate



WES FERNANDES/*NATURE*

Piaget's claim that infants younger than nine months do not understand the permanence of an object that has vanished. When adults view an object disappearing, they tend to show an increase in a particular type of neural oscillation over the right temporal cortex. Johnson, working with colleagues Gergely Csibra and Jordy Kaufman, showed that six-month-old babies show a similar pattern — suggesting that they do keep hidden objects in mind. The same pattern was not observed when the object disintegrated instead of being hidden[4].

Studies such as these have convinced Johnson that babies are not born blank slates, but neither do they possess adult-like concepts about things like number. "My work, I think, goes for a middle ground," he says. He argues that the newborn has basic attention preferences for things such as faces and speech, and that these preferences shape the brain as it develops[5]. Johnson's observation that young babies prefer direct eye contact is one such example; this sets them up to focus on socially relevant parts of their surroundings, which in turn enables them to learn about language and other social cues such as facial expressions.

HAPPY BABY

Working with babies requires specialized kit — particularly for a laboratory that can see as many as 14 in a day. The Babylab kitchen hosts a bottle-warmer, and bathrooms are well stocked with wet-wipes. The waiting room is brightly decorated and scattered with easy-to-clean toys. The laboratories, however, are largely empty and painted a dull battleship grey — a deliberate choice, because babies are easily distracted. "We try to make it as boring as possible, except for the thing we need them to focus on," says Leslie Tucker, coordinator of the Centre for Brain and Cognitive Development, of which the Babylab is part.

Hungry or tired babies do not make for good experiments, so everything is carefully planned around meals and naps. In the waiting room, Caitlin — a four-month-old in stripy blue dungarees — is receiving a last-minute breastfeed before being ushered into a lab. She is participating in a study to assess the development of mimicry in babies: the unconscious tendency of people to frown when someone else frowns, or smile when they smile.

"Mimicry serves important social functions in adults and has even been suggested to be the 'social glue' that binds us together," says Carina de Klerk, who is leading that study at Birkbeck. But very little is known about how, and when, it develops. Some researchers think that it is something babies are born with — newborns have been observed to stick their tongues out in response to an adult doing the same[6]. But "it's not clear if the baby is actually copying, or perhaps they just stick out their tongue whenever something exciting happens", de Klerk says.

She sings to baby Caitlin while sticking electrodes on her temples, cheeks and under her chin. The baby seems unsure, so a research assistant appears, brandishing a garish musical telephone. The art of distraction is a fundamental skill that anyone working in a baby lab must quickly master. "Researchers from other fields come down here and are often horrified at the lack of controls," says Tucker. "You're going to interrupt the experiment if you have to, or make noises to distract them if they look like they're going to cry."

It works: Caitlin is now cooing and smiling. The researchers pause for a moment, while Caitlin's mother takes a photo of her "science baby" on her phone. Then Caitlin is shown a series of video sequences of a woman raising her eyebrows or opening and closing her mouth, interspersed with static pictures of farm animals.

The mimicry experiment is a prime example of the Babylab's mixed-methods approach. Baby Caitlin stares intently at the screen; she does not seem to be copying the woman's actions. But the electrodes on her face may tell a different story: the technique, called electromyography (EMG), picks up electrical activity in her facial muscles, which will indicate if Caitlin is activating her eyebrow area — even if she is not overtly moving it — in response to the woman raising hers. Later in the day, Caitlin is shown the same video sequence while hooked up to NIRS.

NIRS is transforming the ability of researchers to peer into the minds of babies. It was originally adopted by medical physicists at UCL as a technique to help predict the risk of stroke in premature babies. They then began working with Birkbeck researchers to adapt it to answer more fundamental questions[7]. By tracking the flow of oxygenated blood, NIRS allows scientists to see which brain areas become more active in response to external events. For instance, a 2009 study from the Babylab revealed that the brains of five-month-olds already show an adult-like pattern of activation in response to social stimuli, such as a woman playing peek-a-boo with them[8]. In the mimicry study, the researchers want to see if the babies' brains show a similar pattern to those of adults who are mimicking others, which should help to explain if mimicry is partly innate.

But NIRS is not perfect, in part because it cannot measure what is happening in important inner brain regions such as the hippocampus or the amygdala. "The brain is a complex connected circuit. If you only measure a superficial part of that circuit, you can come to the wrong conclusions," Kagan says. To assess these deeper areas, researchers need a technique such as functional magnetic resonance imaging (fMRI),

Not your average lab: the Babylab (left) is designed for infants; a row of EEG 'hairnets' (middle); and an eye-tracking experiment under way (right).

> "You're going to interrupt the experiment if you have to, or make noises to distract them."

which has yielded huge insight into the adult brain. But fMRI is highly sensitive to movement, so babies can be scanned only if they are sedated or asleep, which has severely limited the technique's use.

AN EYE ON AUTISM

Looking time remains an important tool at Birkbeck and elsewhere — although these days, it is assessed not by human observation but by precise eye-tracking technology, such as that being used on baby Ezra. Ezra is a control for the autism and ADHD study: he does not have an older sibling with one of the disorders, so is not considered at high risk. As his attention flits between the apparently random objects on the screen, the reflected infrared light allows psychologist Emily Jones — who directs the project — to gauge precisely what he is looking at, and in which order. "What we tend to find is that typically developing babies will always look first, and longer, at the face, before looking at the other objects," she says.

Autism and ADHD have become a major focus of the Babylab as the prevalence and awareness of the conditions have risen in the past two decades — they are now believed to affect around 4% of the UK population. Last year, in a study of 104 infants, the Birkbeck team showed that infants at high risk of autism were drawn towards the face first, but they seemed to spend less time overall than 'neurotypical' babies in looking at any of the objects — and those that went on to develop autism had the shortest looking time of all[9]. A separate eye-tracking study published by the group earlier this year revealed that nine-month-olds who went on to develop symptoms of autism were more likely to spot the odd-one-out among a group of letters on a screen[10].

It is not completely clear why this is, but the working hypothesis is that these infants are more attentive to the details of what they see, says Teodora Gliga, who led the odd-one-out study. The downside of this could be that children who go on to develop autism find it harder to draw general conclusions about what they are seeing, she says. The study of which Ezra is part aims to extend this work by collecting more-detailed measures from over 400 families — and to identify those features that are strongly associated with the later onset of a developmental disorder. During the five visits that Ezra will make to the Babylab as he grows up, he will be tested using EEG, NIRS and EMG, and his parents will be given extensive questionnaires to assess his language skills, social development, temperament and sleeping patterns.

The team hopes that early brain differences could some day provide indicators — or biomarkers — of autism, which isn't usually diagnosed until close to a child's third birthday. They also hope to find ways to steer brain development back towards a more typical course.

One clinical trial at the Babylab already suggests that early intervention can have an effect. Babies in 28 families with an older sibling with autism were randomly assigned to a group in which they were visited by a therapist at least six times between the ages of seven and ten months, and were compared with a group of high-risk babies who received no therapy. The therapist showed parents videos of them interacting with their child to help understand how their baby was trying to communicate with them, and how to respond. After five months, the team saw hints of improvements in the babies' engagement, attention and social behaviour, compared with controls. But the team acknowledged that many of the results had wide confidence intervals and that it is too early to say whether the intervention will have long-term effects[11].

Johnson hopes that investigations in the toddler lab, when they start, might also eventually find a practical use, helping researchers to devise ways to boost cognitive, attention and memory skills. "I believe we are now at a unique point of convergence between this basic science and the clinical science," he says.

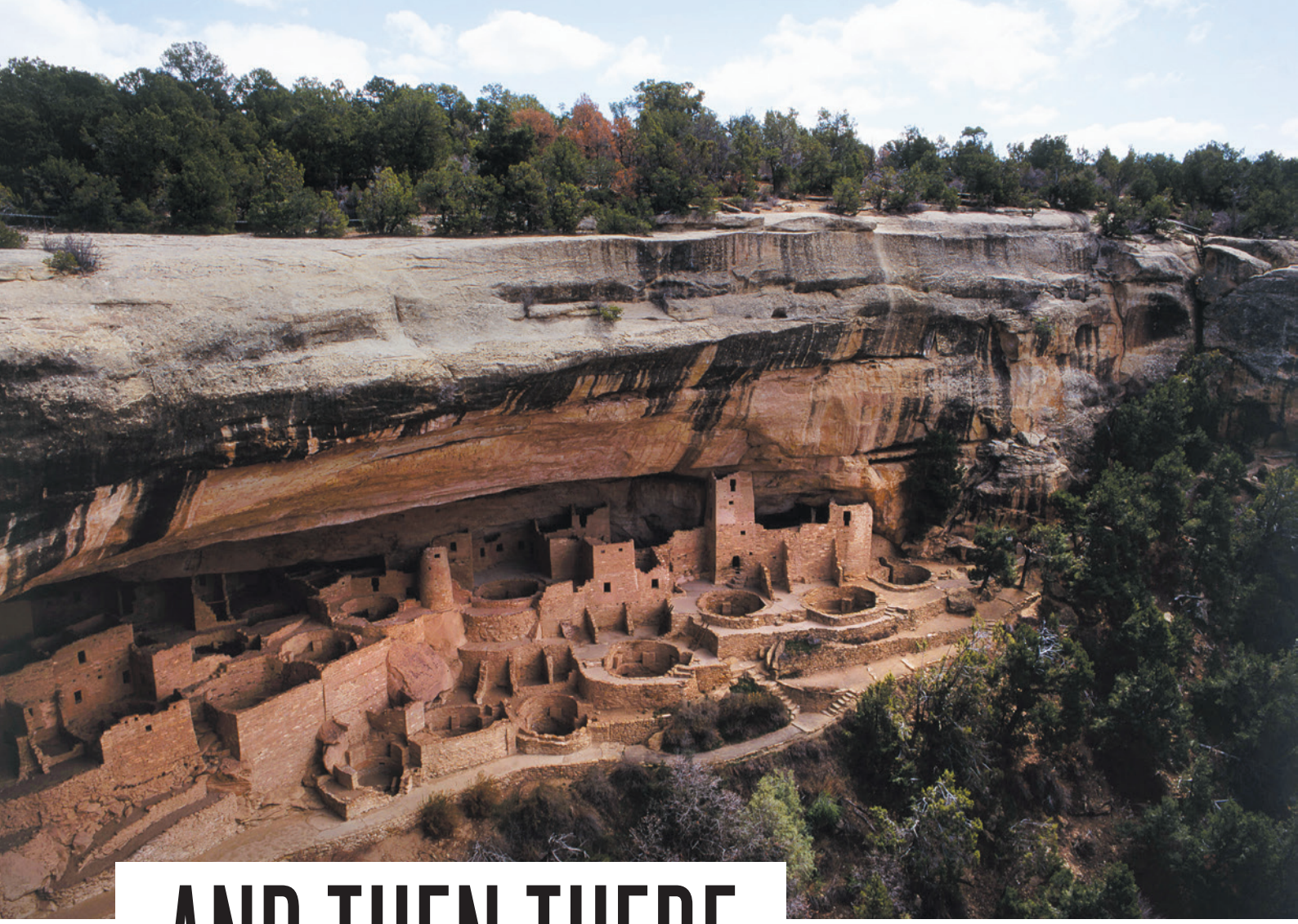
Meanwhile, the techniques continue to evolve. Jones is currently piloting 'gaze-contingent' tasks, which enable babies to become active participants in experiments. "If they can focus their attention on a butterfly flying across the screen, and not get distracted by other things that are happening, then the butterfly keeps flying, so they get rewarded for controlling their attention," Jones says. A more distant goal is to develop ways of using fMRI so that it could be used on awake babies. And there are still so many questions that demand answers. How do differences in the temperaments of babies develop into more complex personality traits as children age? And why can't people remember their earliest months and years?

Baby Ezra will certainly not remember his day in the lab. By late afternoon, his mother is tucking him into the pushchair for his journey home — a 1-hour 45-minute journey to Bristol by train. The trip was worth it, she says, because she was curious to learn what goes on at the Babylab. "I was interested in how Ezra would respond, but also in why those tasks were being done," she says.

Ezra and his mother now have souvenirs of their day: some photos, a certificate of participation and a baby-sized T-shirt. "I'm an infant scientist," it reads. ■

Linda Geddes *is a freelance writer based in Bristol, UK.*

1. Farroni, T., Csibra, G., Simion, F. & Johnson, M. H. *Proc. Natl Acad. Sci. USA* **99,** 9602–9605 (2002).
2. Elsabbagh, M. *et al. Curr. Biol.* **22,** 338–342 (2012).
3. Johnson, M. H. & Horn, G. *Anim. Behav.* **36,** 675–683 (1988).
4. Kaufman, J., Csibra, G. & Johnson, M. H. *Proc. Natl Acad. Sci. USA* **102,** 15271–15274 (2005).
5. Johnson, M. H. *Dev. Cogn. Neurosci.* **1,** 7–21 (2011).
6. Meltzoff, A. N. & Moore, M. K. *Child Dev.* **54,** 702–709 (1983).
7. Blasi, A. *et al. Phys. Med. Biol.* **52,** 6849–6864 (2007).
8. Lloyd-Fox, S. *et al. Child Dev.* **80,** 986–989 (2009).
9. Wass, S. V. *et al. Sci. Rep.* **5,** 8284 (2015).
10. Gliga, T. *et al. Curr. Biol.* **25,** 1727–1730 (2015).
11. Green, J. *et al. Lancet Psychiatry* **2,** 133–140 (2015).



JAMES GRITZ/GETTY

AND THEN THERE WERE NONE

Seven centuries ago, tens of thousands of people mysteriously fled their homes in the American Southwest. Archaeologists are trying to work out why.

BY RICHARD MONASTERSKY

Vultures carve lazy circles in the sky as a stream of tourists marches down a walkway into Colorado's Spruce Canyon. Watching their steps, the visitors file along a series of switchbacks leading to one of the more improbable villages in North America — a warren of living quarters, storage rooms, defensive towers and ceremonial spaces all tucked into a large cleft in the face of a cliff.

When ancient farmers built these structures around the year 1200, they had nothing like the modern machinery that constructed the tourist walkway. Instead, the residents had to haul thousands of tonnes of sandstone blocks, cut timber and other materials down precarious paths to build the settlement, known as Spruce Tree House, in Mesa Verde National Park.

"Why would people live here? That's an important question. It's not an easy place to reach," says Donna Glowacki, an archaeologist now at the University of Notre Dame in Indiana, as she walks among the ruins. Even more perplexing is what happened after they settled there. The villagers occupied their cliffside houses for just a short time before everyone suddenly picked up and left. So did all the other farmers living in the Four Corners region of the American Southwest, where the modern states of Colorado, New Mexico, Utah and Arizona meet (see 'Turbulent times').

All together, nearly 30,000 people disappeared from this area between the mid-1200s and 1285, making it one of the greatest vanishing acts documented in human history. What had been one of the most populous parts of North America became almost instantly a ghost land.

Archaeologists have long puzzled over what drove these farmers, the ancestors of the Pueblo people, from their homes and fields. "That is one of the iconic problems of southwestern — and world — prehistory," says archaeologist Mark Varien, who is executive vice-president of the Crow Canyon Research Institute in Cortez, Colorado. Early scholars blamed

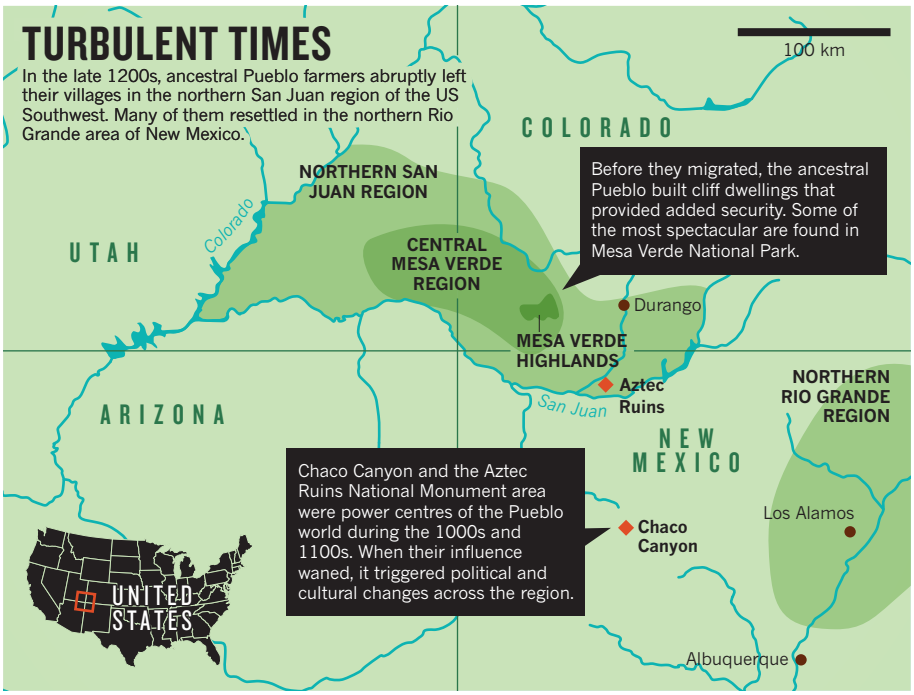
Cliff Palace, a Pueblo dwelling in Mesa Verde National Park, was a thriving village in the 1200s.

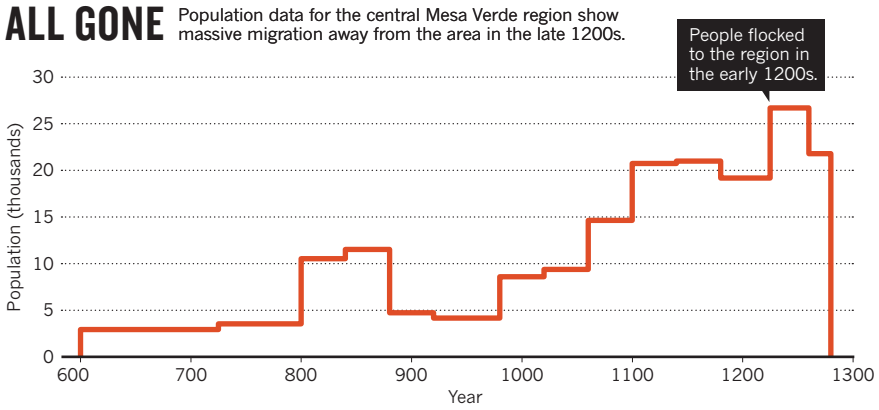
nomads, the ancestors of the Apache and Navajo, for violently displacing the farmers. Over the past couple of decades, the main explanation has shifted to climate — a profound drought and cold snap that hit in the 1270s.

But a series of studies by Glowacki, Varian and other researchers reveals a much more complex answer. The scientists have used detailed archaeological analysis, fine-grained climatic reconstructions and computer models to simulate how ancestral Pueblo families would have responded to their environment. The interdisciplinary strategy has enabled the researchers to examine prehistoric societal changes at a level unattainable in most other regions. "We have enormous detail on this archaeologically. Unparalleled detail," says Steve Lekson, an archaeologist at the University of Colorado Boulder.

The emerging picture is one of a society rocked by troubles until it eventually toppled. More than a century before the Mesa Verde villages emptied out, political disruptions and a monster drought destabilized the entire ancestral Pueblo world. Thousands of people moved into the Mesa Verde region from nearby areas, straining the agricultural capacity of the landscape and eroding established cultural traditions. This led to violent conflicts that further undermined the society, spurring some people to leave. When another drought hit in the late 1200s, the remaining population departed en masse.

Political instability, cultural conflict, violence, overcrowding and drought. Many of the challenges encountered by the ancestral Pueblo seem all too familiar in 2015, as hundreds of thousands of migrants flee from the Middle East and Africa towards Europe. When Glowacki looks at the events of more than seven centuries ago at Spruce Tree House, she sees many similarities. "There was a splintering that went on and an implosion of this political system. It was a rejection, them saying, 'We can't live that way anymore. There has to be a better way'."





SOURCE: S. ORTMAN *J. ARCHAEOL. METH. THEOR.* HTTP://DOI.ORG/8T6 (2014).

STONE WORK

It was chance that first carried Glowacki into the world of the ancestral Pueblo. Before starting graduate school, she ended up in a summer job as a ranger at Mesa Verde National Park, where she fell for the landscape and its archaeology. She has spent the past 23 years, on and off, researching the region's ancient populations.

At Spruce Tree House, Glowacki pulls out a map showing the latest results of an architectural analysis that she is helping the park to carry out. The work is laborious — researchers sometimes sit in front of a wall of sandstone blocks for days, studying the mortar and rocks to work out how the structure was first built and then altered over time.

Gradually, a history of the village has taken shape, showing that people assembled the first set of rooms in the alcove around the year 1200, and added more right up until the last residents abandoned the site around 85 years later. The researchers can narrow construction dates to within a year or two by analysing tree-ring patterns in the wooden support beams in the ceilings and then matching them to an established tree-ring chronology for the region.

Despite the tedious nature of the work, Glowacki says that it never loses its appeal. "There are rooms that are fully intact, and you can stand in them — and they were built in the 1240s. In this country, being able to stand in something that was built at that time is really pretty magical."

The cliff dwellings were a last resort for the park's prehistoric Pueblo residents. When farmers first arrived in the region around AD 600, they settled on the fertile highlands above the canyons, which gave them easier access to their fields. But by 1200, something began to force them over the edge into the giant alcoves that naturally form in the sandstone cliffs.

Insights into that shift are emerging thanks to a major interdisciplinary effort called the Village Ecodynamics Project (VEP), which launched in 2002. Funded by the US National Science Foundation, the nearly US$2.5-million initiative is assessing how social and environmental factors influenced the populations of prehistoric Pueblo farmers from about 600 to 1300, says Tim Kohler, the VEP's principal investigator and an archaeologist at Washington State University in Pullman.

In one strand of research, the team drew on the rich history of archaeology in the region to compile a database of 18,000 prehistoric sites, which allowed them to measure the population and how it shifted over time[1]. With such a massive database, the researchers could look at population changes in narrow time bands averaging about 40 years (see 'All gone').

"There are not many places in the world where archaeologists can look at changes in such discrete slices of time," says Varien, who is a co-principal investigator of the VEP. The analysis[1] suggested that people started leaving the Mesa Verde region at least 15 years before

the drought hit. "It looks as though the final depopulation began with a trickle and ended with a flood," says Scott Ortman, an archaeologist at the University of Colorado Boulder who developed the model for the project's population analysis.

Another part of the VEP looked at how the farmers fed themselves. The researchers used temperature and precipitation estimates from tree-ring data to create a model of where the communities could have grown maize (corn) each year, which was their main source of food. The calculations of this 'maize niche' did a good job of explaining how many people settled in different regions, says Kohler.

The team's latest data show that when growing conditions improved, the population density spiked, more than doubling in some regions. But one place defied that pattern: Mesa Verde National Park. When farming became easier, people actually moved out of that area. And, paradoxically, when times grew tough, more people moved in.

Kohler and his colleagues suggest that these movement patterns have to do with topography. The park stands higher than the surrounding landscape, so it gets more precipitation. And because the highlands tilt to the south, cold air drains off, leaving Mesa Verde warmer than the surrounding lowlands. So when the region faced drought or a cold spell, farmers congregated in the more-reliable Mesa Verde area — something researchers had not appreciated before now, says Kohler. "People have been working in this area for 100 years, and I don't think they ever realized it," he says of such a climate pattern.

VIRTUAL REALITY

The VEP researchers have also conjured up a virtual version of the past. The team constructed a computer model of the landscape and then seeded it with households that could grow maize, hunt, collect water and wood and move to new sites if they failed to secure enough resources. By comparing the simulations to the archaeological record, the researchers can examine factors that might have driven ancient populations to migrate. "It's really a new way of doing archaeology," says Varien.

Kohler says that he sometimes switches on the graphics during a simulation to watch the behaviour of the dots that represent households. Scattered randomly at first, they scurry around until their inhabitants can harvest enough resources. Then, they form into settlements, which grow rapidly to a point when they can no longer sustain themselves — and so the households move again. But there is a limit to how much Kohler can watch. "Even on modern, fast processors, when the agents get into the thousands, it slows down and it's no longer fun," he says.

By comparing the simulations to the actual population data, the researchers discovered[2] some interesting discrepancies during the 1100s and 1200s. In the model, the farmers spread out farther across the landscape than they actually did in reality. So something seems to have caused the real ancestral Pueblo to huddle together more tightly than expected.

Kohler and his colleagues wondered whether fear might have been a factor. To find out, they surveyed the archaeological literature and tracked levels of violence in the area through time by tallying how many skeletons had broken arm bones, fractured skulls or other signs consistent with acts of aggression. Some had apparently died in massacres, and there was even evidence of cannibalism at certain sites.

Between 600 and 1000, the Mesa Verde region was relatively peaceful, but rates of violence rose in the mid-1000s and spiked again in the late 1200s, right before the ancient Pueblo left, the researchers reported last year[3]. "What we found was that people were more clumped up than the model predicted precisely in times when there was a lot of violence on the landscape," says Kohler.

There has been some scepticism among archaeologists about the use of agent-based modelling, but Kohler says that it has been useful in this case: the inconsistency between the simulations and the real data prompted the researchers to look at violence in a new way. "That disjunction identifies for us interesting questions," he says.

processing, so the team will graduate to a supercomputer for future simulations, which are planned for later this year or early next year. Nothing of this scale has been done before in the field, says Kohler. "Archaeologists do not have the reputation of being users of high-performance computing environments," he says. "But I don't think we'll be the end of the road for this kind of work."

Among the ruins at Spruce Tree House, Glowacki takes a different approach. As a collaborator on the VEP project, she does not discount the importance of drought and short growing seasons. But she focuses on some of the other factors that also stressed the ancestral Pueblo society.

The signs are in the houses that fill the Spruce Canyon alcove. The architectural-documentation project has taught Glowacki that the residents there updated their homes just as much as people in New York or London today. "Even when they were living there, they were making changes and adding walls and doors and doing all of this remodelling."

CULTURE CLASH

Some of these alterations point to dramatic events. In the mid-1200s, structures associated with one of the founding families were burned: fire damage can be seen in one room and in a kiva, a circular depression that served as the family's ceremonial space. The fire does not seem to be accidental, Glowacki says. Rather, it could have been part of a ritual changeover in ownership or it might reflect someone forcing out one of the original clans. "At the very least, that suggests there were some significant changes in the clans or families that were using the structures — or in part of the leadership there."

> "IT GOT REALLY BAD AND REALLY NASTY, AND THEY WANTED TO GET AWAY FROM IT."

Most researchers think that the majority of violent acts occurred within ancestral Pueblo communities: one village attacking another over food resources or neighbours turning on each other. More than half the skeletons from some periods bore signs of trauma. "They are one of the most violent societies we've ever studied," says Kohler.

But not all of their troubles came from within. Some unusual-looking projectile points have turned up at massacre sites that date to just before the Pueblo people left the Mesa Verde region, so invading nomads might have had a role in forcing the farmers from their homes.

In the next stage of the VEP project, researchers plan to look at how food shortages might have contributed to violence. The new version of the agent-based model is more sophisticated than the last, allowing households to form social groups that compete with each other for access to agricultural lands. Leaders can emerge, fighting can erupt between groups and people can migrate away from Mesa Verde to an area farther south in New Mexico, where many ancestral Pueblo are thought to have resettled.

This all amounts to a huge step up in

Other rooms in the alcove were also burned, including a tower that may have served as a defensive structure. Taken together, the architectural evidence provides a detailed view of friction in the village, she says. "There was some sort of conflict and people left, presumably, and new people came in and remade these spaces."

Around the Pueblo region, there are many signs of cultural change leading up to and during the 1200s. Glowacki, along with some other archaeologists, thinks that such adjustments had to do with shifting political allegiances in that part of the world.

During the mid-1000s and early 1100s, the centre of power among the Pueblo people was located about 150 kilometres south of the Mesa Verde area, in New Mexico's Chaco Canyon.

ROBERT JENSEN/MESA VERDE NATL PARK

In Spruce Tree House, a ladder leads down into a sunken ceremonial space known as a kiva.

In the 1100s, an extension of the Chaco political order rose up at a site now called Aztec Ruins National Monument, midway to Mesa Verde. The Chaco–Aztec culture was socially stratified, with massive residences in which the elites lived. Smaller versions of the elite 'great houses' have been found in villages to the north, which reveals the broad influence of the Chaco–Aztec political order.

Then, an awful drought between 1130 and 1150 apparently weakened that order, and new types of practice emerged. In the Mesa Verde region, some communities built more-inclusive spaces, such as open plazas, and they removed the roofs from some large kivas, allowing broader participation in rituals[4].

The changes in public and ceremonial spaces demonstrate the waning influence of the Chaco–Aztec polity, which had previously unified the Pueblo world. "What is happening is you have this dissolution and splintering," Glowacki says. That may have contributed to the increased violence and served to drive farmers from their highland villages towards the more-secure alcoves along the cliff faces.

These political upheavals may also partially explain why people started to abandon the Mesa Verde area decades before the drought of the mid-1270s hit. The combination of political instability, social upheaval and then a rotten climate was too much to take, she says. "It got really bad and really nasty, and they wanted to get away from it."

Kohler sees parallels with the collapse of the classic Mayan civilization in the ninth century, as well as with events in the Middle East today. In the case of the Mesa Verde exodus, researchers can look in detail not only at why and when people left, but also at what happened afterwards. "We need to understand migration streams better," he says. "We have the advantage of the long view."

FINDING PEACE

Whatever forced the Pueblo to uproot themselves, tens of thousands of people left the Four Corners region in search of something better. And many apparently found what they were looking for. When the exodus began, the ancestral Pueblo migrated in several different directions: some to the southwest into Arizona and some to southern New Mexico. Archaeologists have long suspected that many settled along the Rio Grande river in northern New Mexico, a couple of hundred kilometres southeast of the Mesa Verde region. That hypothesis is supported by population data, which show that the Rio Grande region became more crowded; VEP studies[5] have indicated that between 1250 and 1300, the population in this area swelled from 8,000 to 18,000 people. By the early decades of the 1300s, it was close to 25,000, Ortman says.

When they settled in their new home, the Mesa Verde people made a clear break from their former lives. Analysis by Kohler, Ortman and their colleagues[3] shows that rates of violence were much lower than before. And the Pueblo made social changes as well. "The migrants do not appear to be trying to continue with the society and traditions of the Four Corners. They were trying to leave them behind," says Ortman. The Pueblo villages that grew up after 1300 reflect a much more communal type of society, in which multiple families shared kivas and residents gathered in open ceremonial spaces.

There was also a political change, says Lekson, who has studied the elite residences at Chaco Canyon and Aztec Ruins. "They shucked off all the nobles and the kings, and they never had them again. They figured out how to run villages without that apparatus."

Even today, southwestern Pueblo villages continue to embrace an egalitarian society. Ortman finds inspiration in the evolution of Pueblo culture after the collapse. "Pueblo people had to create those values and institutions that reflect them as a result of their past struggles," he says.

And that system has been remarkably successful. Pueblo villages have retained their culture and languages to a much stronger degree than most other Native American communities, he says. "Some of the Pueblos that emerged after the Mesa Verde migration have been able to withstand 500 years of European colonization," says Ortman. "One could say that those communities have weathered European colonization better than almost any other society in the world — certainly within the United States."

At Spruce Tree House, Glowacki has seen how strong those traditions still are. Just a few weeks earlier, she took part in a workshop that included some teachers who are Pueblo and who demonstrated how they grind maize. Even that mundane chore took on spiritual dimensions as the teachers made offerings to their ancestors who once inhabited the cliff dwelling. To the modern Pueblo, the centuries-old structures are not abandoned ruins but still echo with the spirits of those who came before.

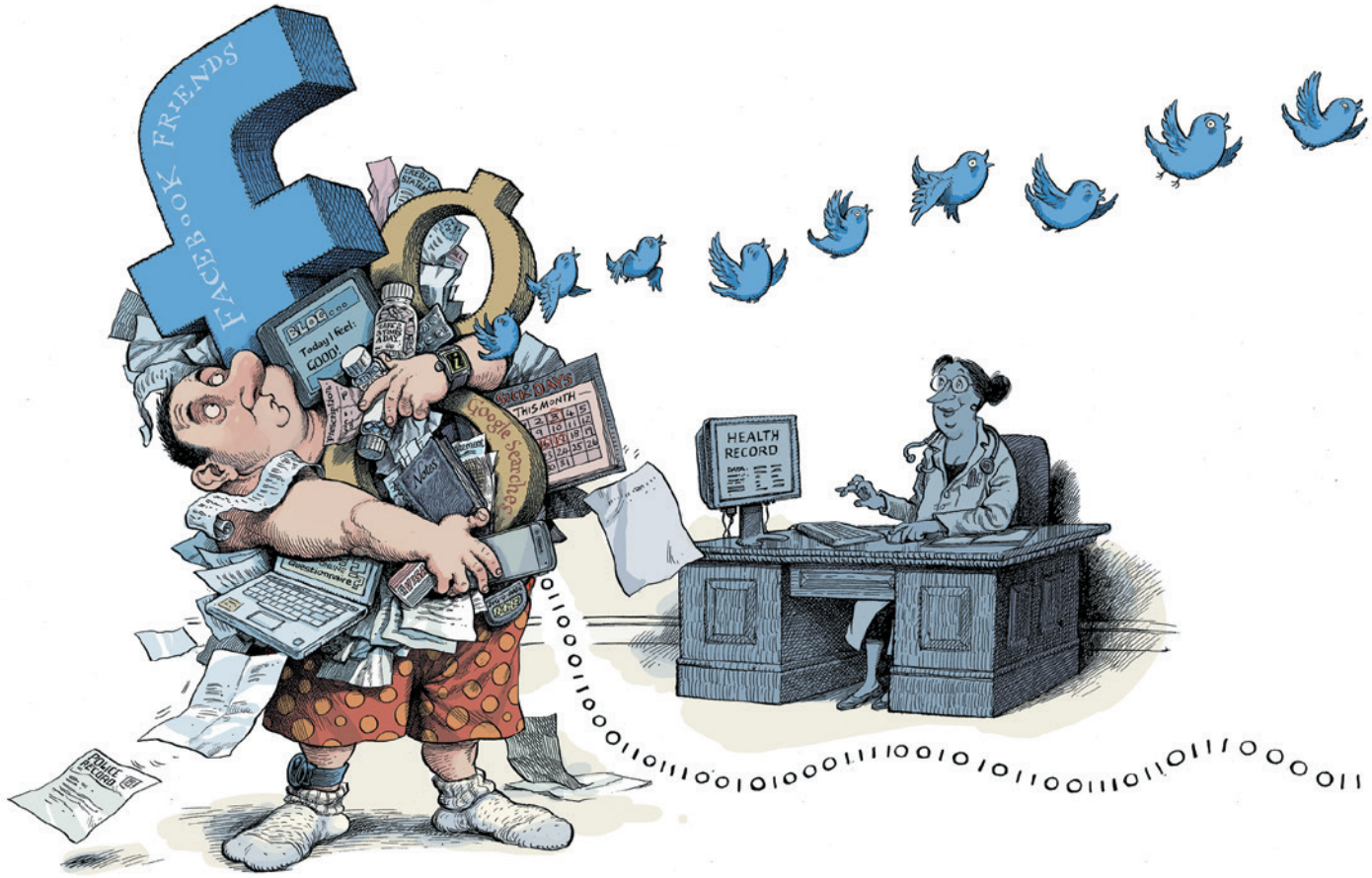
"It was a really beautiful moment," says Glowacki. "What I think makes Pueblo culture really interesting and perhaps unique is the long arc of Pueblo history. There's a lot we can learn about how a society faces really difficult times, adversities — and fundamentally reorganizes and transforms their culture." ■

Richard Monastersky *is a features editor for* Nature.

1. Schwindt, D. M. *et al. Am. Ant.* (in the press).
2. Kohler, T. A. & Varien, M. D. (eds) *Emergence and Collapse of Early Villages: Models of Central Mesa Verde Archaeology* (Univ. California, Berkeley Press, 2012).
3. Kohler, T., Ortman, S., Grundtisch, K., Fitzpatrick, C. & Cole, S. *Am. Antiq.* **79,** 444–464 (2014).
4. Glowacki, D. M. *Living and Leaving* (Univ. Arizona Press, 2015).
5. Ortman, S. G. *Winds from the North* (Univ. Utah Press, 2012).

COMMENT

ILLUSTRATIONS BY DAVID PARKINS



Make sense of health data

Develop the science of data synthesis to join up the myriad varieties of health information, insist **Julian H. Elliott**, **Jeremy Grimshaw** and colleagues.

If you are wondering whether exposure to some chemical could increase your chances of getting colon cancer, you could easily find supportive evidence from animal experiments. You might then discover that epidemiological studies tell a different story.

There have never been more options when it comes to measuring factors relevant to health. We can sequence our entire genomes and those of our bacteria, viruses and tumours. In principle, every visit to the doctor can be tracked from electronic medical records. Information on physiology, behaviours, diets, movements and interactions with others can be extracted from wearable devices, smartphone apps and social-networking sites[1]. And thanks to the open-access movement and a shift in data-sharing norms, more data are being made publicly available.

Yet sifting through the information to find answers to questions about health is becoming increasingly difficult, even for the experts. The data exist in disparate domains, are generated using different methods, and are stored in different infrastructures — from the private servers of hospitals to global platforms, such as dbGaP, an open database of genotypes and clinical information.

POOLING DATA

We believe that to consolidate data from different sources into comprehensive and coherent bodies of evidence on which decision-makers can act, researchers need to better exploit current methods and tools for data synthesis — and to develop superior ones.

Researchers usually try to obtain insights by pooling the same kind of data, such as ▶

from clinical trials. But because different study and data types tend to have distinct strengths and weaknesses, a much richer understanding can emerge when different kinds of information are combined.

The drug cisapride, for instance, was licensed in the United States in 1993 to treat heartburn, on the basis of data collected in clinical trials over ten years. Yet the drug's association with fatal heart-rhythm disturbances[2] was understood only when data from clinical trials were consolidated with those from large, long-term cohort studies, which recorded cisapride's effects in thousands of people.

Likewise, the picture obtained from conventional influenza surveillance (which involves collecting data from primary-care clinics) can lag behind what is actually happening on the ground. Google collects real-time information based on the use of search terms related to flu symptoms, but these findings can be inaccurate. The best insights almost certainly come from aggregating these different data types[3].

So how can we bring together the multiple, extremely diverse data sets that are now becoming available?

Formal methods for 'evidence synthesis' — in which multiple sources of data are combined to obtain new insights — were first developed in the social sciences in the 1970s. The techniques have since been adapted in many branches of science, and they underpin high-impact decision-making, for example in drug licensing[4]. They generally involve identifying and collating all the available and relevant data; assessing each data source's strengths and vulnerability to bias; and deciding how to handle the different sources of data depending on their rigour and the question being asked (some data may be excluded, for instance). Then, if appropriate, a meta-analysis or qualitative assessment can be conducted, incorporating the information[5].

For example, a UK group combined[6] data from clinical trials with those from cohort studies in a meta-analysis to assess the effectiveness of anti-D, a drug given to some pregnant women to prevent them from producing antibodies against their babies. In this case, potential sources of bias, such as different clinics providing care for the women in cohort studies, were systematically identified, and their impact was minimized.

Yet many researchers immersed in the combination and analysis of large data sets that are vulnerable to spurious correlations, such as genomic or electronic-medical-record data, are unaware of evidence-synthesis tools and their potential usefulness. Conversely, many experts in evidence synthesis are unfamiliar with the methods often used to analyse large data sets relevant to health.

We believe that the core elements of evidence synthesis must be combined with other data sciences to develop new ways to make sense of diverse data.

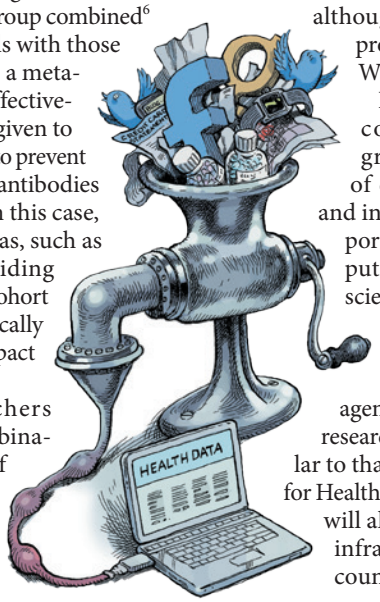
MANAGING BIAS

Scientists need to work out why, when and how to combine diverse data — for instance, should physical-activity data from clinical records, online questionnaires and wearable devices be combined? As well as addressing when and how to combine diverse individual-level data, scientists need to grasp the risks of bias associated with each data type and incorporate such risks into their analyses. For clinical trials and observational studies of the effects of interventions, analysts can use the Cochrane Risk of Bias approach. Similar methods are needed to enable the detection and reduction of bias in other data types, such as social-networking and mobile-phone data.

"Society does not need more islands of data analysis."

Also needed are agreed ways to capture and represent information on potential sources of bias. Organizations investing in infrastructure and standards for health data, such as Health-Level 7, need to incorporate this layer of metadata (data about data) into their systems.

Methods to deal with bias must be incorporated into new analytical systems developed to guide decision-making in health care — including those based on natural-language processing and machine learning. Transparent and independent evaluations of these new systems will also be important, although challenging to achieve for proprietary systems such as IBM Watson.



In the short to medium term, conferences, funding programmes and a restructuring of departments in universities and institutes will be crucial to support collaborations between computational biologists, computer scientists, clinical and population-health researchers and specialists in evidence synthesis. For instance, major granting agencies should invest in dedicated research-methods programmes similar to that of the UK National Institute for Health Research. Targeted investment will also be needed to develop data infrastructure in poor regions and countries. In the long term, a new type of analyst, adept at appraising and combining diverse data types appropriately, may emerge.

JOINING THE DOTS

What could these shifts mean in practice? One of the aims of the US Precision Medicine Initiative (PMI) is to prevent people from getting cancer. This means understanding the effects of myriad genomic, behavioural and environmental factors and their interactions. The value of the initiative will be enhanced if data from these very different domains can be combined appropriately and easily.

Another aim of the initiative is to develop new cancer therapies. Better systems for data synthesis would inform drug development with richer and more accurate insights from the 'omics' sciences, animal studies and early human trials. Moreover, health-care funders such as Britain's National Health Service and Medicare in the United States could better understand a drug's benefits and harms in the real world by synthesizing data from clinical trials, cohort studies, patient experiences reported through mobile and social applications, and drug-surveillance systems. (These include the US Sentinel Initiative and the Canadian Network for Observational Drug Effect Studies, which pool data from different health-care systems to monitor the adverse effects of licensed drugs.)

We are not proposing a one-model-fits-all approach. But society does not need more islands of data analysis that support conflicting inferences. As large and diverse data sets become ever more plentiful, we must ensure that rigorous and trustworthy methods to make sense of the data are developed in parallel. ■

Julian H. Elliott *is senior research fellow at the Australasian Cochrane Centre at Monash University, and head of clinical research in the Infectious Diseases Unit at Alfred Hospital, Melbourne, Australia.* **Jeremy Grimshaw** *is senior scientist at Ottawa Hospital Research Institute and professor of medicine at the University of Ottawa, Canada.* **Russ Altman, Lisa Bero, Steven N. Goodman, David Henry, Malcolm Macleod, David Tovey, Peter Tugwell, Howard White, Ida Sim.**
e-mail: julian.elliott@alfred.org.au

1. Weber, G. M., Mandl, K. D. & Kohane, I. S. *J. Am. Med. Assoc.* **311,** 2479–2480 (2014).
2. Wysowski, D. K. & Bacsanyi, J. *N. Engl. J. Med.* **335,** 290–291 (1996).
3. Lazer, D., Kennedy, R., King, G. & Vespignani, A. *Science* **343,** 1203–1205 (2014).
4. Institute of Medicine. *Finding What Works in Health Care: Standards for Systematic Reviews* (National Academies Press, 2011).
5. Chalmers, I. *Ann. Am. Acad. Pol. Soc. Sci.* **589,** 22–40 (2003).
6. Turner, R. M. *et al. PLoS ONE* **7,** e30711 (2012).

The authors declare competing financial interests. For details, and for full author affiliations, see go.nature.com/scxwp9.

IMAGINECHINA/CORBIS

Many choices that people consider their own are already determined by algorithms.

Build digital democracy

Open sharing of data that are collected with smart devices would empower citizens and create jobs, say **Dirk Helbing** and **Evangelos Pournaras**.

Fridges, coffee machines, toothbrushes, phones and smart devices are all now equipped with communicating sensors. In ten years, 150 billion 'things' will connect with each other and with billions of people. The 'Internet of Things' will generate data volumes that double every 12 hours rather than every 12 months, as is the case now.

Blinded by information, we need 'digital sunglasses'. Whoever builds the filters to monetize this information determines what we see — Google and Facebook, for example. Many choices that people consider their own are already determined by algorithms. Such remote control weakens responsible, self-determined decision-making and thus society too.

The European Court of Justice's ruling on 6 October that countries and companies must comply with European data-protection laws when transferring data outside the European Union demonstrates that a new digital paradigm is overdue. To ensure that no government, company or person with sole control of digital filters can manipulate our decisions, we need information systems that are transparent, trustworthy and user-controlled. Each of us must be able to choose, modify and build our own tools for winnowing information.

With this in mind, our research team at the Swiss Federal Institute of Technology in Zurich (ETH Zurich), alongside international partners, has started to create a distributed, privacy-preserving 'digital nervous system' called Nervousnet. Nervousnet uses the sensor networks that make up the Internet of Things, including those in smartphones, to measure the world around us and to build a collective 'data commons'. The many challenges ahead will be best solved using an open, participatory platform, an approach that has proved successful for projects such as Wikipedia and the open-source operating system Linux.

A WISE KING?

The science of human decision-making is far from understood. Yet our habits, routines and social interactions are surprisingly predictable. Our behaviour is increasingly steered by personalized advertisements and search results, recommendation systems and emotion-tracking technologies. Thousands of pieces of metadata have been collected about every one of us (see go.nature.com/stoqsu). Companies and governments can increasingly manipulate our decisions, behaviour and feelings[1].

Many policymakers believe that personal data may be used to 'nudge' people to make healthier and environmentally friendly decisions. Yet the same technology may also promote nationalism, fuel hate against minorities or skew election outcomes[2] if ethical scrutiny, transparency and democratic control are lacking — as they are in most private companies and institutions that use 'big data'. The combination of nudging with big data about everyone's behaviour, feelings and interests ('big nudging', if you will) could eventually create close to totalitarian power.

Countries have long experimented with using data to run their societies. In the 1970s, Chilean President Salvador Allende created

computer networks to optimize industrial productivity[3]. Today, Singapore considers itself a data-driven 'social laboratory'[4] and other countries seem keen to copy this model.

The Chinese government has begun rating the behaviour of its citizens[5]. Loans, jobs and travel visas will depend on an individual's 'citizen score', their web history and political opinion. Meanwhile, Baidu — the Chinese equivalent of Google — is joining forces with the military for the 'China brain project', using 'deep learning' artificial-intelligence algorithms to predict the behaviour of people on the basis of their Internet activity[6].

The intentions may be good: it is hoped that big data can improve governance by overcoming irrationality and partisan interests. But the situation also evokes the warning of the eighteenth-century philosopher Immanuel Kant, that the "sovereign acting … to make the people happy according to his notions … becomes a despot". It is for this reason that the US Declaration of Independence emphasizes the pursuit of happiness of individuals.

Ruling like a 'benevolent dictator' or 'wise king' cannot work because there is no way to determine a single metric or goal that a leader should maximize. Should it be gross domestic product per capita or sustainability, power or peace, average life span or happiness, or something else?

Better is pluralism. It hedges risks, promotes innovation, collective intelligence and well-being. Approaching complex problems from varied perspectives also helps people to cope with rare and extreme events that are costly for society — such as natural disasters, blackouts or financial meltdowns.

Centralized, top-down control of data has various flaws. First, it will inevitably become corrupted or hacked by extremists or criminals. Second, owing to limitations in data-transmission rates and processing power, top-down solutions often fail to address local needs. Third, manipulating the search for information and intervening in individual choices undermines 'collective intelligence'[7]. Fourth, personalized information creates 'filter bubbles'[8]. People are exposed less to other opinions, which can increase polarization and conflict[9].

Fifth, reducing pluralism is as bad as losing biodiversity, because our economies and societies are like ecosystems with millions of interdependencies. Historically, a reduction in diversity has often led to political instability, collapse or war. Finally, by altering the cultural cues that guide peoples' decisions, everyday decision-making is disrupted, which undermines rather than bolsters social stability and order.

Big data should be used to solve the world's problems, not for illegitimate manipulation. But the assumption that 'more data equals more knowledge, power and success' does not hold. Although we have never had so much information, we face ever more global threats, including climate change, unstable peace and socio-economic fragility, and political satisfaction is low worldwide. About 50% of today's jobs will be lost in the next two decades as computers and robots take over tasks. But will we see the macro-economic benefits that would justify such large-scale 'creative destruction'? And how can we reinvent half of our economy?

The digital revolution will mainly benefit countries that achieve a 'win–win–win' situation for business, politics and citizens alike[10]. To mobilize the ideas, skills and resources of all, we must build information systems capable of bringing diverse knowledge and ideas together. Online deliberation platforms and reconfigurable networks of smart human minds and artificially intelligent systems can now be used to produce collective intelligence that can cope with the diverse and complex challenges surrounding us.

"Big data should be used to solve the world's problems."

A DIGITAL NERVOUS SYSTEM

The Nervousnet project is working on this. It began as a tool for scientists to experiment with the Internet of Things. For example, social interactions can be studied by anonymously tracing the physical proximity of people (given their informed consent).

Nervousnet now enables anyone to measure and analyse aspects of the world in real time. The Nervousnet app allows users to activate or deactivate about ten smartphone sensors that measure, for example, acceleration, light and noise. A range of other functions are being shaped by the core research and development team at ETH Zurich and about a dozen research groups in Europe, Japan and the United States. The project is funded by the European Commission, Delft University of Technology in the Netherlands and philanthropists. It is also supported by volunteer developers. We aim for global collaboration and benefits, even if there will be different variants in the end (as happened for Unix operating systems, for example).

Unlike initiatives for the Internet of Things spearheaded by big technology companies, Nervousnet is run as a 'citizen web', built and managed by its users. Inspired by Wikipedia and OpenStreetMap, people can interact with Nervousnet in three ways. They can contribute data, analyse the crowdsourced data sets, and share code and ideas. Anyone can create data-driven services and products using a generic programming interface. The aim is to yield societal benefits, business opportunities and jobs.

Several Internet of Things platforms and data-science projects share Nervousnet's vision; none has its scope. They focus on participatory data collection; decentralized communication services; or big-data analytics. Nervousnet is designed to meet all three objectives. It will also enable real-time measurement and feedback to support self-organizing systems. For example, self-controlled traffic lights responding to local vehicle flows can reduce urban congestion and outperform today's centralized systems.

Nervousnet uses distributed data storage and distributed control, so that it is resilient to attacks and centralized manipulation attempts, easy to scale up, and tolerant to faults. Because data encryption is not enough, a secure personal-data store will be needed to allow each user to determine which data to share with whom, and for what purpose.

Attracting users is a challenge. We will be adding elements of gaming to make participation more enjoyable, as well as a micro-payment system to reward and incentivize digital co-creation. Because critics may worry about the responsible use of bottom-up systems, Nervousnet will integrate reputation systems, qualification mechanisms and self-governance by community moderators.

In the long run, measurements tailored to specific purposes and a combination of crowdsourced data generation, curation and analysis will outperform the currently fashionable big-data analytics approach. Just as the open standards of the World Wide Web created unprecedented opportunities and a multibillion-dollar economy, the right framework for the Internet of Things and digital society could foster an age of prosperity. ■

Dirk Helbing *is professor, and* **Evangelos Pournaras** *is a postdoctoral researcher, in computational social science at the Swiss Federal Institute of Technology in Zurich, Switzerland.*
e-mail: dhelbing@ethz.ch

1. Kramer, A. D. I., Guillory, J. E. & Hancock, J. T. *Proc. Natl Acad. Sci. USA* **111,** 8788–8790 (2014).
2. Epstein, R. & Robertson, R. E. *Proc. Natl Acad. Sci. USA* **112,** E4512–E4521 (2015).
3. Medina, E. *Cybernetic Revolutionaries* (MIT Press, 2011).
4. Harris, S. 'The Social Laboratory' *Foreign Policy* (29 July 2014); available at http://go.nature.com/k1xd9n
5. Storm, D. 'ACLU: Orwellian Citizen Score, China's credit score system, is a warning for Americans' *Computerworld* (7 October 2015); available at http://go.nature.com/3pq8b4
6. Hsu, C.-P. 'Baidu welcomes China's military to join China Brain project on AI systems' *Want China Times* (2015); available at http://go.nature.com/myhd1x
7. Page, S. E. *The Difference: How the Power of Diversity Creates Better Groups, Firms, Schools, and Societies* (Princeton Univ. Press, 2008).
8. Pariser, E. *The Filter Bubble: What the Internet Is Hiding from You* (Viking/Penguin, 2011).
9. Andris, C. *et al. PLoS ONE* **10,** e0123507 (2015).
10. Helbing, D. *The Automation of Society Is Next* (in the press); preprint available at http://go.nature.com/b1gnkx

PETTY OFFICER 1ST CLASS SHANE T. MCCOY/US NAVY/GETTY

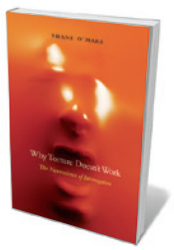
Detainees are held at the United States' Guantanamo Bay detention camp in Cuba in 2002.

NEUROSCIENCE

Tortured reasoning

Lasana T. Harris commends a book exposing the lack of scientific basis to 'enhanced interrogation techniques'.

In 2009, following the abuse of prisoners at its Guantanamo Bay detention camp, the US government made a significant decision. It moved the responsibility for 'enhanced interrogation techniques' from the CIA to a new government organization: the High-Value Detainee Interrogation Group (HIG). The move upset many CIA insiders; torture had been in their toolkit since the early days of the cold war. The remarks of one official at a HIG-organized conference on torture in Washington DC can be summed up as: how could a new agency, created to both conduct and study torture, replace the decades of practice and perfection attained by the CIA? By adding a scientific component, responded the newly appointed head of the HIG.



Why Torture Doesn't Work: The Neuroscience of Interrogation
SHANE O'MARA
Harvard University Press: 2015.

This exchange highlights the theme of neuroscientist Shane O'Mara's *Why Torture Doesn't Work*. Rightly, O'Mara takes a moral stand against torture (forced retrieval of information from the memories of the unwilling). However, instead of simply providing utilitarian arguments, he argues that there is no evidence from psychology or neuroscience for many of the specious justifications of torture as an information-gathering tool. Providing an abundance of gruesome detail, O'Mara marshals vast, useful information about the effects of such practices on the brain and the body.

For instance, he explains why, physiologically, it is ludicrous to claim that stress, pain and fear will coerce a suspect to surrender critical information. The prolonged release of stress hormones such as cortisol damages the hippocampus — a brain structure crucial for encoding and retrieving memories — as well as the prefrontal cortex, which is implicated in decision-making and executive control processes. Such damage works in opposition to the goal of torture. Furthermore, chronic stress creates a negative feedback loop, causing enlargement and hyperresponsiveness of the amygdala, the brain structure that underlies emotional salience, directs attention, enables learning and communicates with most of the brain.

Another striking example that O'Mara discusses is the effect on the brain of sleep deprivation. The practice was described in the 'Torture Memos' — legal memoranda drafted in 2002 by US deputy assistant attorney general John Yoo, advising the CIA and President George W. Bush on the use of torture. Officially limited to a maximum of 180 hours, and often combined with physical restraint, isolation, starvation and beatings, sleep deprivation has been used to coerce subjects into revealing information.

The memos further argue that sleep deprivation is harmless. O'Mara, however, discusses research suggesting that it erodes memory processes and general cognitive function by flooding the brain with glucocorticoid hormones. Even military scientists have produced literature that admits psychophysiological issues with sleep deprivation. In 1990, Paul Naitoh and his colleagues at the US Naval Health Research Center in San Diego, California, published evidence that the practice leads to an increase in circulating stress hormones and the development of psychomotor epileptic discharges (P. Naitoh *et al. Occup. Med.* **5,** 209–237; 1990). They argued, too, that if combined with other stressors, such as food and water deprivation and waterboarding, sleep deprivation could negatively affect respiratory–cardiovascular function.

Yet some officials and politicians continue to make announcements that run counter to such scientific evidence. Former Pennsylvania senator and Republican presidential hopeful Rick Santorum, for instance, commented in a 2011 interview that after being broken, people become cooperative. Most shocking may be this year's revelation that a handful of officials in the American Psychological Association were complicit in torture by the United States after the September 2001 attacks on New York and the Pentagon, thus providing a veil of scientific legitimacy to the practice.

Torture also affects the torturer. The cognitive dissonance required to inflict suffering results in symptoms similar to those of post-traumatic stress disorder, O'Mara warns. He cites Joshua Phillips's *None of Us Were Like This Before* (Verso, 2010), which describes how many US veterans who had engaged in torture in Iraq experienced intense guilt or turned to substance abuse once back in the United States. Interviews with former interrogators in Northern Ireland, published by Ian Cobain in *Cruel Britannia* (Portobello, 2012), reveal that many believed what they had done was wrong, but saw it as a desperate attempt to end the violence engulfing their society.

Given that information obtained under torture is rarely reliable (because the victim will generally say anything to make ▶

the pain stop) O'Mara recommends an alternative: conversation. Having a conversation with a detainee may yield results comparable, and probably superior, to those obtained from torture. He cites three pieces of evidence.

First is a 1993 study by Stephen Moston and Terry Engelberg of police interrogations, which found that of more than 1,000 detainees, only 5% refused to talk (S. Moston and T. Engelberg *Polic. Soc.* **3,** 223–237; 1993). Second, research by Robin Dunbar and his colleagues finds that 40% of what we reveal in conversation is related to the self, suggesting that refusing to self-disclose is very difficult (R. I. M. Dunbar *et al. Hum. Nat.* **8,** 231–246; 1997). Third, a study by Diana Tamir and Jason Mitchell showed that people are willing to forgo money to talk to others about themselves. Indeed, the nucleus accumbens (part of the brain's reward circuitry) activates during such an opportunity, suggesting that people find disclosure intrinsically rewarding (D. I. Tamir and J. P. Mitchell *Proc. Natl Acad. Sci. USA* **109,** 8038–8043; 2012). O'Mara does acknowledge that the difficulties of having such a conversation with a non-compliant person demand advanced social skills that are comparable to those of clinical psychologists and psychiatrists, who often deal with non-compliant patients. He suggests that alternative approaches such as virtual reality and role playing may be useful for information gathering during interrogation.

> ***"Conversation with a detainee may yield results comparable, and probably superior, to those obtained from torture."***

Why then, given its uselessness in eliciting valuable information, do people torture? It is a form of vengeance or punishment, intended to discourage the victim from future transgressions and to communicate to others that harm will not be tolerated. In some cases, it occurs because the torturer believes that terrorists have mental illnesses. In science, however, punishment is not a viable response to someone with such an illness — just as torture is not a viable method for gathering information, as O'Mara repeatedly points out. ■

Lasana T. Harris *is a senior lecturer in experimental psychology at University College London, and a guest lecturer in social and organizational psychology at Leiden University in the Netherlands. He studies the neuroscience of dehumanization and prejudice.*
e-mail: lasana.harris@ucl.ac.uk

Steve Jobs (Michael Fassbender) confronts his daughter Lisa (Perla Haney-Jardine) in *Steve Jobs*.

FRANÇOIS DUHAMEL

INNOVATION

A binary life

A polished biopic of tech titan Steve Jobs fails to plumb fully his inner contradictions, finds **Timo Hannay**.

Steve Jobs
WRITTEN BY AARON SORKIN
DIRECTED BY DANNY BOYLE
Universal: 2015.

The closest I came to meeting Steve Jobs was in the late 2000s, shortly after the birth of the iPhone. I was attending Foo Camp, a California mustering of digital demigods. Jeff Bezos of Amazon was a regular; the year before, Google co-founder Larry Page had turned up in his helicopter. Everyone but me took such things in their stride. That year, however, there was something different in the air: a rumour had spread that Steve Jobs himself might join us. He never showed up, but such was his unique status that even his absence generated more excitement than the presence of other tech giants.

Blessed as he was with formidable taste and rock-star showmanship, Jobs was always going to stand out from the crowd of awkward nerds (like me) who populate much of the technology landscape. Add to this his death at the height of his powers, and we have all the ingredients of a legend. This is not undeserved. Many technologists talk of changing the world; Jobs actually did so. More than anyone else, he broke down the barriers between technology and humanity, helping to turn computers into consumer products. Then, with the iPhone, he pulled off the reverse, turning an established consumer product into a computer.

How best to understand such a life? Jobs's answer was to invite high-flying writer and former media executive Walter Isaacson to pen his biography — a superb account published within days of Jobs's death. *Steve Jobs* (Simon and Schuster, 2011) is likely to remain the closest we will ever get to a definitive account.

The film version of Isaacson's blockbuster is a highly competent creation — as you would expect from writer Aaron Sorkin (*The Social Network*, *The West Wing*) and director Danny Boyle (*Slumdog Millionaire*, *Trainspotting*). The dialogue zips along at 100 beats per minute; the acting (especially by Michael Fassbender in

the title role) is at times outstanding; and the direction is as slick as that of any other Hollywood offering. Yet many people will watch this film to better understand its subject — and by that measure, it falls short.

The plot hinges on Jobs's relationship with his daughter, Lisa Brennan-Jobs, and plays on the contrast between his lavishing of obsessive attention on his latest electronic brainchild and his ignoring, or disowning, of his flesh and blood. It does this by going backstage at three seminal product launches: those of the Macintosh in 1984, the NeXT Computer in 1988 and the iMac in 1998. This convenient three-act structure, which catches Jobs at three key moments in his life, also serves as a metaphor for the contrast between his suave public persona and his chaotic life.

This leaves a lot out. And therein lies the main weakness of this film: there are umpteen other contradictions to explore in Jobs. He was simultaneously a hippy and a control freak. He was an ascetic drawn to mysticism who built the world's preeminent consumer-products company. He was egocentric and impossible, inspiring both incredible feats of engineering (starting with the design of the Apple II by co-founder Steve Wozniak) and deep affection (despite frequently taking credit for the work of Wozniak and others).

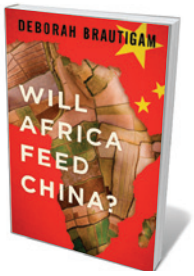
Of course, covering all this ground in a two-hour film would be difficult. But the setting means that Jobs's close colleagues, relatives and key antagonists must all be at the launches with him, wanting to discuss their gripes in the same few minutes before he is due to step on stage. (In one amusing 'meta' moment, Fassbender actually notes precisely this.) This frequently stretches credibility too far.

The first two-thirds of the film thus struggle to engage — and will probably confuse people unfamiliar with the story and the cast of characters. It includes plenty of wonderfully quotable lines and aphorisms from the book, such as Jobs's burning desire to "put a dent in the universe". But the rat-a-tat-tat form feels more like a collage than a coherent narrative. In short, it could have done with a dose of Jobsian minimalism. That said, the film redeems itself in the third act — rather like Jobs's career.

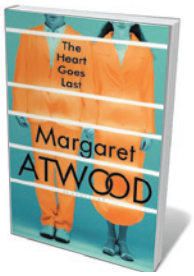
If you want an impressionistic, almost dreamlike montage of key moments in Jobs's life, see *Steve Jobs*. If you want to understand Jobs the man, you will be disappointed. But see the film anyway: it makes a great trailer for the book. ■

Timo Hannay *is the founder of SchoolDash, an education data analytics firm based in London.*
e-mail: timo@hannay.net

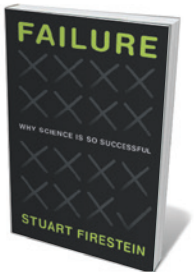
Books in brief



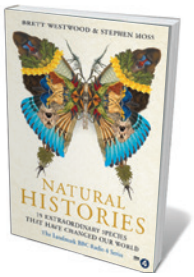
Will Africa Feed China?
Deborah Brautigam OXFORD UNIVERSITY PRESS *(2015)*
Starting in 2008, China — with more than 20% of the global population and just 9% of the arable land — was said to be buying up swathes of African farmland. In her cogent analysis, international-development specialist Deborah Brautigam cuts her own swathe through myths about this relationship. She marshals fresh case studies to reveal that Chinese companies own just 250,000 hectares of African land, while the country has no government policy on overseas farming. Far from being the first ripple of an imperial storm, she argues, Chinese interests in Africa largely follow in Western footsteps.



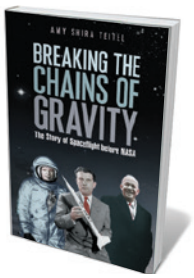
The Heart Goes Last: A Novel
Margaret Atwood BLOOMSBURY *(2015)*
Stan and Charmaine struggle to survive in a squalid, lawless near future. The Positron Project, a social experiment in which they spend alternating stints in prison and suburbia, seems to offer a way out — at first. Doyenne of speculative fiction Margaret Atwood is on grimly hilarious form here as tour guide to a macabre society given over to unregulated science, social cleansing, identity loss and profiteering. She prods satirically at issues from industrial farming (headless-chicken production aimed at "meat growth efficiencies") to sexbots, and even fits in a subplot featuring a horde of Elvis impersonators.



Failure: Why Science Is So Successful
Stuart Firestein OXFORD UNIVERSITY PRESS *(2015)*
Biologist Stuart Firestein's energetic sequel to *Ignorance* (Oxford University Press, 2012) explores the centrality of failure in the scientific endeavour. Naturalist Ernst Haeckel's erroneous ideas about ontogeny and phylogeny, for instance, helped to spawn the field of embryology. Firestein ranges widely, looking at failure in contexts ranging from pharma to funding. At base, however, this is a close examination of how repeated failure refines problems, clarifying the way forward — a challenge that in turn sparks the courage and clarity of mind needed for incisive investigation.



Natural Histories: 25 Extraordinary Species That Have Changed our World
Brett Westwood and Stephen Moss JOHN MURRAY *(2015)*
Based on an eponymous BBC Radio 4 series, this collaboration with London's Natural History Museum explores the biology and cultural histories of selected flora and fauna. Naturalist Brett Westwood and writer Stephen Moss present an idiosyncratic list, including mandrill, oak, coral, cockroach and whale. Out of myriad gripping stories, their take on the lion resonates: the imposing beast may be a cultural ubiquity, yet African populations have diminished catastrophically from 400,000 in 1950 to fewer than 30,000 today.



Breaking the Chains of Gravity: The Story of Spaceflight before NASA
Amy Shira Teitel BLOOMSBURY SIGMA *(2015)*
In this straightforward chronicle, science journalist Amy Shira Teitel traces NASA's 'prequel'. However familiar, the early discoveries of rocketeers such as Romanian physicist Hermann Oberth still thrill, as does (in a very different way) the crucial input of former Nazi and rocket designer Wernher von Braun. Teitel delivers on detail, such as the exploits of supersonic-flight pioneer Chuck Yeager; but the whole needs more synthesis and never quite soars. **Barbara Kiser**

Correspondence

Volkswagen and the road to Paris

In the wake of the Volkswagen emissions-testing scandal (see *Nature* http://doi.org/723; 2015), this month's climate summit in Paris needs to roll out an international framework for regulating emissions — with strong incentives and tough penalties. Voluntary national measures are no longer enough.

Volkswagen's gaming of emissions testing underscores the urgency of reinventing transport. Governments must implement electric transport systems and plan for combustion-free inner cities, by expanding such schemes as London's ultralow-emission zone, due in 2020. Chancellor Angela Merkel could fast-track zero-emission zones for Berlin, Hamburg, Munich and Frankfurt by 2025, for example — restoring Germany's environmental lead.

The Volkswagen debacle should be treated as an Enron moment for sustainability measurement and valuation, with a comparable overhaul of the requirements for corporate accounting and evaluation. Programmes such as the Redefining Value initiative of the World Business Council for Sustainable Development can capture environmental externalities, including impacts on climate, biodiversity and health. Now we just need governments to incorporate these into their regulatory and stock-market requirements.

Gail Whiteman, Harry Hoster *Lancaster University, UK. g.whiteman@lancaster.ac.uk*

DEFRA responds to badger-cull critique

In calculating the effectiveness of the latest UK badger-culling targets for controlling bovine tuberculosis, Christl Donnelly and Rosie Woodroffe do not consider the uncertainties in estimating badger populations or how information collected during culling is used to evaluate the success of culls in real time (*Nature* **526,** 640; 2015).

Experience shows that there is greater uncertainty associated with badger population estimates than previously thought. A post-cull assessment by the Department for Environment, Food and Rural Affairs (DEFRA), using the number of badgers removed and reductions in sett occupancy, suggests that badger abundance may have been overestimated. Using the mean of the population estimate to establish a minimum number to be culled, as implied in Donnelly and Woodroffe's calculations, leads to a high probability of a culling objective that could greatly exceed actual badger numbers.

The current culls use methods similar to a trial that ran from 1998 to 2006 in southwest England and the west Midlands. The trial achieved a roughly 70% reduction in badgers, with large variance between trial zones, based on post-hoc assessments. Applying similar culling effort to a zone should converge on a similar outcome to the trial.

To reduce the badger population by a similar proportion as in the trial while providing an achievable objective, the government has set an initial minimum culling number at the lower end of the estimated population range. Information gathered during the cull will be used to assess whether this number should be increased. The most up-to-date data about the badger population are used to assess whether the culls are removing enough badgers and are therefore likely to achieve a similar outcome to the trial.

Comparison with control zones, where there has been no culling, provides no indication that culling has increased disease in cattle, as was widely predicted in advance of the culls (see go.nature.com/grk4ri).

Ian L. Boyd *University of St Andrews, UK; and DEFRA, UK. ilb@st-andrews.ac.uk*

China emissions: stop subsidizing emitters

China needs to resolve its conflicting policies on reducing carbon emissions and on increasing economic growth if it is to implement a cap-and-trade system successfully (see *Nature* **526,** 13–14; 2015).

For example, the government subsidizes several industries that are big energy consumers and generate excessive emissions and pollution. China's coal-driven iron and steel industry is one such case, despite its overcapacity, low profit and vicious competition.

By June 2015, six months after China's revised Environmental Protection Law came into effect, 2,556 listed companies were in receipt of government subsidies that totalled 250 times more than the fines for environmental damage (see go.nature.com/ozprpc (in Chinese) and D. Liu *Nature* **525,** 321; 2015).

These absurd subsidies hamper the transformation of industry to cleaner production and distort resource allocation through local protectionism and lobbying. They should be backed by firmer legislation or abolished.

Xin Miao *Harbin Institute of Technology, Harbin, China. xin.miao@aliyun.com*

China emissions: alter energy markets

China has issued a nationwide cap-and-trade programme and a series of laws to cut its carbon emissions by 40–45% between 2005 and 2020 (see *Nature* **526,** 13–14; 2015 and G. Wagner *et al. Nature* **525,** 27–29; 2015). The realities of running such complicated systems and pricing schemes are daunting, however.

Obstacles include promotion of local government officials, which depends not on how well they protect the environment but on how they help to develop the economy. And more commercial incentives are needed for China to implement ways to reduce emissions.

Although the government has vowed to make the energy sector more accountable in market terms, administrative interventions continue to be the norm. The energy market is dominated by monopolies, and prices are tightly controlled by the administration. These problems must be addressed if China is to use its resources efficiently.

Dayuan Li, Shenggang Ren *Business School of Central South University, Hunan, China.* **Xiaohong Chen** *Hunan University of Commerce, China. bigolee@163.com*

Europe's first '3Rs' governmental centre

In September, the German government opened a nationwide centre at the Federal Institute for Risk Assessment that is legally committed to protecting animals used for scientific purposes. The initiative is the first of its kind in Europe and is scientifically independent of executive and political advisory bodies. It aims to encourage greater transparency and raise standards of animal welfare by adopting an interdisciplinary approach.

Known as Bf3R (www.bf3r.de), it will encourage European research to meet the '3Rs' targets for animal experimentation (for replacement, reduction and refinement; see go.nature.com/yidbm2). It will lead the way in enforcing the country's Animal Welfare Act and European Directive 2010/63/EU on the protection of lab animals.

Bf3R will also advise on legal and other requirements, helping authorities and researchers across Europe to communicate proper animal-protection practice to other scientists and to the public.

Gilbert Schönfelder *Federal Institute for Risk Assessment (BfR); and Charité — University Medicine Berlin, Germany.* **Barbara Grune, Andreas Hensel** *BfR, Germany. gilbert.schoenfelder@bfr.bund.de*

SUSTAINABILITY

Australia at the crossroads

A modelling study argues that comprehensive policy change could limit Australia's environmental pollution while maintaining a materials-intensive path to economic growth. But other paths are worth considering. SEE ARTICLE P.49

BENJAMIN L. BODIRSKY & ALEXANDER POPP

Despite Australia's vastness and its swathes of untouched nature, its per-capita environmental footprint is one of the biggest worldwide. Because it is a major exporter of agricultural products, coal and other emissions-intensive commodities, there is great concern that binding climate agreements could harm the country's economy. In 2014, under then prime minister Tony Abbott, the current conservative government replaced a carbon-tax policy with inefficient mitigation subsidies[1]. Abbott was toppled from the party leadership in September 2015. His successor, Malcolm Turnbull, was once a strong proponent of a carbon-trading scheme, but it remains uncertain whether environmental policies will be reformed under his leadership.

On page 49 of this issue, Hatfield-Dodds *et al.*[2] argue that Australia can stick to its materials-intensive industries and enjoy continued high economic growth while reducing its impacts on climate, water and biodiversity. The authors show that greenhouse-gas emissions can be mitigated through efficiency improvements in production processes, and even more through carbon removal by planting forests (afforestation) and carbon capture and storage. The premise in any case is a comprehensive pricing of emissions.

Hatfield-Dodds and colleagues' assessment, the most comprehensive conducted for Australia so far, is based on the Australian National Outlook 2015, a report[3] prepared by the country's Commonwealth Scientific and Industrial Research Organisation. The authors used nine linked simulation models to estimate the performance of Australia's economy in a global market, with a particular focus on the agriculture, energy and transport sectors, which exert the largest environmental pressures on land, water and climate. The modelling framework is exemplary in bridging scales between global, national and sub-national dynamics. This cross-scale approach could, and should, become seminal for future regional assessments.

The study produces 20 scenarios for Australia's future, exploring possible domestic developments in regard to lifestyle, policy and technological progress. All scenarios are embedded in one of four possible settings for global change, characterized by different population trajectories and by different global carbon prices, leading to 2, 3 or 6 °C of global warming above pre-industrial levels in the year 2100. The authors' models then provide projections, under each scenario, for rates of technology adoption in the energy, transport and agricultural sectors; for production, income, and trade; and for environmental indicators such as water usage, land clearing and greenhouse-gas emissions.

The findings indicate that Australia's gross domestic product will more than double by 2050 in all scenarios. However, without carbon pricing, greenhouse-gas emissions would increase by up to 90% in the same period. Even with a carbon tax at a similar level to that in force in 2012–14, Australia's emissions are projected to rise by about 25% by 2050. Complying with a 2 °C global-warming target will require higher taxes, which Hatfield-Dodds *et al.* show can be reached most cost-effectively

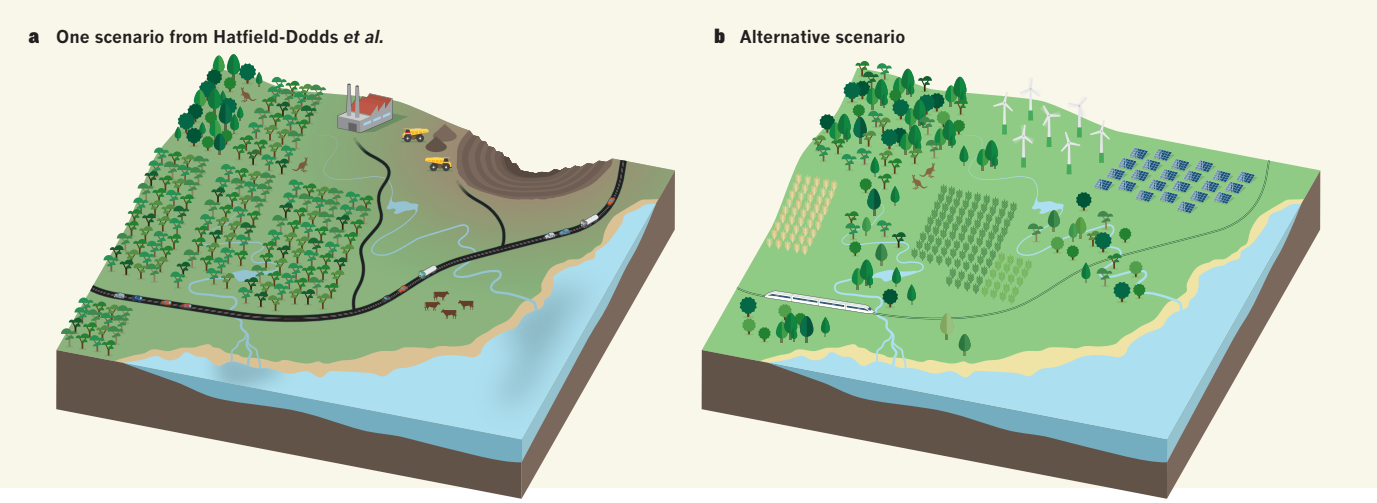


Figure 1 | Possible paths. **a**, Scenario modelling presented by Hatfield-Dodds *et al.*[2] suggests that Australia could maintain its economic growth and its typically materials-intensive lifestyles, while reducing its environmental impacts. Under this scenario, fossil fuels continue to be burned, but in combination with carbon capture and storage. The transport sector switches to electric and hybrid cars. Agriculture is intensified and dominated by forest plantations to sequester carbon, while biodiversity reserves and seawater desalination produce ecosystem services. **b**, An alternative pathway, not simulated by the authors, is a structural change towards a labour- and technology-intensive economy, with dematerialized lifestyles. Energy is obtained from renewable sources and public transport is expanded. Agriculture gradually shifts from resource-intensive livestock and feed production towards diverse high-value horticulture, and natural and agricultural systems are integrated. We suggest that this pathway would be more resilient to technological or institutional failure.

in Australia through large-scale afforestation and renaturation programmes. In an international carbon market, such greening programmes can become a profitable export industry through the sale of carbon credits.

The general outcome of this Australian assessment is in line with the findings of the *Special Report on Emissions Scenarios* produced by the Intergovernmental Panel on Climate Change (IPCC)[4], which concluded that immediate and global action to limit warming to 2 °C by 2100, in combination with the full availability of key technologies, would entail losses in global consumption of 2–6% (median 3.4%) in 2050 and 3–11% (median 4.8%) in 2100. But Hatfield-Dodds and colleagues' regional study argues that even Australia, with its high dependence on fossil-fuel and agricultural exports, and with high per-capita emissions, does not need to fear increased mitigation costs, because it can remain one of the most cost-efficient producers.

However, although the study shows that Australia can reduce emissions and environmental impact while sticking to its materials-intensive production and consumption patterns, the authors assess only a selection of potential pathways (Fig. 1). Within the literature on future scenarios[4–6], the possibilities considered by Hatfield-Dodds *et al.* describe a rather optimistic future in terms of political institutions and technological performance, and envisage a society open to trade and migration and with materialistic lifestyles. Ecosystem services are valued, but with a curative rather than a preventive approach to environmental damage. Focusing on this strand of scenarios might mask certain risks and opportunities.

One such risk is that future technologies will perform less well than we expect them to. For example, the performance of carbon-capture-and-storage technologies and of large-scale afforestation enormously influence the challenges and mitigation costs of reaching ambitious climate targets[7]. In a world that relies on resource-intensive growth, if such mitigation options fail, this could escalate abatement costs or render climate targets unachievable.

Society might also fail to establish the institutional framework required to embed a materials- and energy-intensive economy into environmental systems. Such a framework requires not only a timely international agreement on global carbon pricing, but also the regulation of other indirect environmental costs that are not reflected by market prices (externalities), such as groundwater use or nutrient pollution. Hatfield-Dodds and colleagues' study clearly shows that, without such policy frameworks, problems rapidly emerge — for example, fast-growing forests planted for carbon sequestration can lead to extreme water scarcity in certain catchment areas. Other side effects could include the increased use of pesticides and fertilizers when afforestation reduces the areas available for crops[8], or the disruption of marine ecosystems as a result of water desalination[9].

The study convincingly argues that lifestyle changes, such as reduced working time, are not sufficient to solve environmental problems. But such changes do help to relieve pressure in the water–energy–food–climate–biodiversity nexus[10] and might lessen the grave consequences of technological or institutional failure. Even in high-abatement scenarios, Hatfield-Dodds and colleagues estimate that per-capita energy demand will not fall below current levels, and that the global demand for animal products will double. Here, they may underestimate the potential for behavioural change, which was also highlighted in the IPCC's Fifth Assessment Report[7].

This work reinforces the appraisal that global pricing of greenhouse gases is essential to mitigate climate change effectively and efficiently[7], and that it should be supported by a general regulation of environmental externalities to avoid unwanted effects. Anchoring mitigation commitments in a global climate treaty has the capacity to protect Australia's economy from unfair competition and to allow continued growth.

Beyond this, this paper and other findings of the Australian National Outlook[3] should trigger debate on how to shape Australia's future. Continuous, resource-intensive growth is one possible pathway, but it will require powerful institutions to restrain the pressure on environmental systems. Another pathway could be an economy shaped by technology and labour instead of energy and resources, allowing less-strict regulation to keep the economy within environmental boundaries. The structural change needed for the latter pathway could be initiated by investing carbon-tax revenues in education and science, establishing markets for flexible electricity consumption, providing bicycle and public-transport infrastructure and promoting healthy and sustainable diets. Australia is free to choose which path to follow. ■

Benjamin L. Bodirsky *and* **Alexander Popp**
are at the Potsdam Institute for Climate Impact Research, 14412 Potsdam, Germany.
B.L.B. *is also at the Commonwealth Scientific and Industrial Research Organisation, St Lucia, Australia.*
e-mails: bodirsky@pik-potsdam.de; popp@pik-potsdam.de

1. Schiermeier, Q. *Nature* **511,** 392–392 (2014).
2. Hatfield-Dodds, S. *et al. Nature* **527,** 49–53 (2015).
3. Hatfield-Dodds, S. *et al. CSIRO Australian National Outlook: Living Standards, Resource Use, Environmental Performance and Economic Activity, 1970–2050*; www.csiro.au/nationaloutlook (CSIRO, 2015).
4. IPCC. *Emissions Scenarios* (eds Nakićenović, N. *et al.*) (Cambridge Univ. Press, 2000).
5. Millennium Ecosystem Assessment. *Ecosystems and Human Well-being: A Framework for Assessment* (Island, 2003).
6. O'Neill, B. C. *et al. Glob. Environ. Change* http://dx.doi.org/10.1016/j.gloenvcha.2015.01.004 (2015).
7. IPCC. *Climate Change 2014: Mitigation of Climate Change* (eds Edenhofer, O. *et al.*) (Cambridge Univ. Press, 2014).
8. Bodirsky, B. L. & Müller, C. *Environ. Res. Lett.* **9,** 111005 (2014).
9. Becker, N., Lavee, D. & Katz, D. *J. Water Resource Protect.* **2,** 1042–1056 (2010).
10. Smith, P. *et al. Glob. Change Biol.* **19,** 2285–2302 (2013).

MATERIALS SCIENCE

Droplets leap into action

What could cause a water droplet to start bouncing on a surface? It seems that a combination of evaporation and a highly water-repellent surface induces droplet bouncing when ambient pressure is reduced. SEE LETTER P.82

DORIS VOLLMER & HANS-JÜRGEN BUTT

On page 82 of this issue, Schutzius *et al.*[1] report a remarkable phenomenon: at low pressure, droplets of water resting on an extremely water-repellent surface spontaneously jump and bounce. In some cases, the height of each bounce increases, like a gymnast jumping on a trampoline. The findings add to our understanding of how droplet–surface interactions can prevent the accumulation of water or ice on surfaces.

Ice accretion on surfaces is a big problem in cold regions, particularly for aviation, shipping or offshore industries[2]. Strategies to minimize ice adhesion include using either smooth or highly water-repellent (superhydrophobic) surfaces. Superhydrophobic surfaces are covered with tiny protrusions that have low interfacial energy, which minimizes their attraction to liquids.

A water or ice droplet resting on a superhydrophobic surface sits on top of the protrusions, so that the main part of the droplet's underside is separated from the surface's substrate by a thin layer of air[3] (Fig. 1). The small contact area between the water or ice and the protrusions ensures low ice adhesion. However, the remaining adhesion is usually still sufficiently strong to keep ice in place.

Furthermore, because the volume of water increases during freezing, water droplets can expand into the space between protrusions upon freezing, increasing both the contact area and the adhesion of the resulting ice.

So how can low pressure cause droplets on a superhydrophobic surface to start trampolining? Schutzius and co-workers propose that two effects need to be considered. First, as noted above, the surface reduces the droplets' adhesion. Such low adhesion has been shown to cause droplet jumping when two droplets merge, because the adhesion energy is easily overcome by the surface energy that is released by the merging[4] (surface energy quantifies the disruption of intermolecular bonds that occurs in a liquid when a surface is formed). The second effect is evaporation. When water evaporates in still air, the rate of evaporation is limited by the ability of the water vapour to diffuse. Reducing the pressure of the surrounding gas increases the diffusion, and thus the rate of evaporation.

In Schutzius and colleagues' study, gas and water vapour are rapidly pumped away from the experimental chamber. A film of water vapour therefore remains only in the gap between the droplets' underside and the surface substrate, because the water vapour's escape from this region is inefficient. An overpressure therefore builds up in the gap — that is, the pressure in the gap becomes higher than that of the surrounding atmosphere.

The authors argue that the droplet jumps once the force induced by the overpressure on the droplet overcomes gravity and adhesion. The gravitational force on droplets of 1 millimetre radius is about ten times lower than the adhesive force, so less than 10% of the total overpressure needed to cause jumping is used to overcome gravity. But, gravity is, of course, required for the droplet to fall back to the surface.

When droplets land back on the surface, they spread and their kinetic energy is transformed into surface energy. This spreading is followed by retraction into an almost spherical droplet, during which the surface energy is transformed back into kinetic energy and the droplet bounces up again. For millimetre-sized droplets, spreading and retraction take several milliseconds[5,6].

By calculating the volume of water vapour that can pass through the gap between the underside of the droplet and the surface's substrate per unit of time, the authors show that overpressure builds up beneath the droplet about ten times faster than the typical contact time of a droplet with the surface. The overpressure induces an upward force on the droplet that adds to the force caused by the conversion of surface energy to kinetic energy when the droplet retracts. This additional force can increase the height of the droplet's bounces, until a maximum height is reached after a few rebounds.

At sufficiently low ambient pressure, the temperature in the droplet can fall below its freezing point because of cooling caused by evaporation[7]. Schutzius *et al.* report that jumping can also be triggered by freezing of such supercooled water droplets — the latent heat released on freezing causes a sudden overpressure and the droplet jumps off the substrate.

Droplet trampolining resembles the Leidenfrost phenomenon, which can be observed when water is spilt on a hot pan. A liquid droplet in close contact with a hot, solid surface gives rise to a vapour layer beneath the droplet; this vapour keeps the liquid from making direct physical contact with the surface. Typically, the droplet immediately starts to hover and move around. By contrast, the onset of trampolining can be fine-tuned by adjusting the time at which the system is depressurized. Another difference is that the Leidenfrost effect is caused by an imposed temperature difference between the droplet and surface, whereas droplet trampolining is caused by a pressure difference generated by the droplet itself.

Inertia and viscous dissipation (the conversion of a fluid's surface and kinetic energy into internal energy) typically dominate the rebound of a droplet from a superhydrophobic surface. By contrast, trampolining results from a uniformly increasing force acting on the droplet's lower surface.

A force also acts on a droplet's lower surface during pancake bouncing[8] — a phenomenon that occurs when droplets collide with superhydrophobic surfaces made from an array of submillimetre-spaced, tapered protrusions. During pancake bouncing, droplets hitting the surface penetrate substantially into the array, whereby kinetic energy is transferred to interfacial energy. This process is followed by upward motion of the droplet out of the array through capillary action, during which the interfacial energy is transformed back into kinetic energy. The droplets then bounce off the surface in a pancake-like shape.

Both the trampolining and pancake-bouncing mechanisms reduce the contact time of bouncing droplets compared with bouncing on a normal surface[6]. However, unlike pancake bouncing, droplet trampolining is expected to occur for a large variety of surface topographies, as long as the gap beneath the droplet is kept thin (at least 100 times less than the droplet diameter at a pressure of about 0.05 bar). If the gap is too large, water vapour would escape too quickly to have an effect and the overpressure in the textured surface would not be high enough.

Although Schutzius and colleagues' observations are fascinating, reducing atmospheric pressure is not a practical way of preventing icing in outdoor areas. And even for smaller areas, much energy is consumed in reducing the ambient pressure. Furthermore, evaporation eventually causes the droplets to become so small that they come to rest — although bouncing has not been maintained indefinitely in any other drop-impact experiments.

Nevertheless, the authors have vividly illustrated that simple experiments can yield surprising results. Applying underpressure to a system is the most common way to enhance evaporation, and is often used in chemical and technical laboratories. Who would have guessed that it could produce such spectacular dynamics? ■

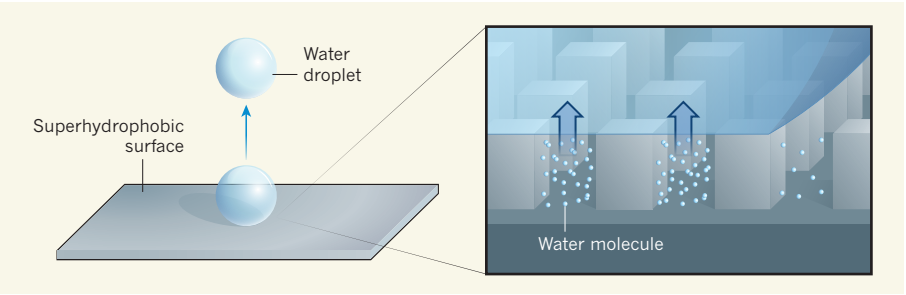


Figure 1 | Mechanism of droplet trampolining. Schutzius *et al.*[1] report that, in a low-pressure environment, water or ice droplets placed on superhydrophobic surfaces (which are covered with micrometre-sized hydrophobic protrusions) can spontaneously jump and bounce. The authors propose that, when a droplet is in contact with the surface, water-vapour molecules from the droplet escape more slowly from the gap beneath the droplet's underside than they do from elsewhere. The pressure in the gap therefore becomes larger than ambient pressure, generating a force (arrows) that lifts the droplet up. (Inset adapted from ref. 9.)

Doris Vollmer *and* **Hans-Jürgen Butt**
are at the Max Planck Institute for Polymer Research, Mainz 55128, Germany.
e-mails: vollmerd@mpip-mainz.mpg.de; butt@mpip-mainz.mpg.de

1. Schutzius, T. M. *et al. Nature* **527,** 82–85 (2015).
2. Lv, J., Song, Y., Jiang, L. & Wang, J. *ACS Nano* **8,** 3152–3169 (2014).
3. Bormashenko, E. Yu. *Wetting of Real Surfaces* (De Gruyter, 2013).
4. Boreyko, J. B. & Chen, C.-H. *Phys. Rev. Lett.* **103,** 184501 (2009).
5. Richard, D., Clanet, C. & Quéré, D. *Nature* **417,** 811 (2002).
6. Bird, J. C., Dhiman, R., Kwon, H.-M. & Varanasi, K. K. *Nature* **503,** 385–388 (2013).
7. Jung, S., Tiwari, M. K. & Poulikakos, D. *Proc. Natl Acad. Sci. USA* **109,** 16073–16078 (2012).
8. Liu, Y. *et al. Nature Phys.* **10,** 515–519 (2014).
9. Butt, H.-J. *et al. Curr. Opin. Colloid Interface Sci.* **19,** 343–354 (2014).

METABOLISM

Light on leptin link to lipolysis

Cutting-edge experiments show that the hormone leptin, which is secreted by fat cells, promotes fat loss by activating the release of catecholamine signalling molecules from neurons wrapped around the fat cells.

JOHAN RUUD & JENS C. BRÜNING

Anyone wanting to lose a few extra pounds might well wish that fat could be burnt at the flick of a switch. As Zeng *et al.*[1] report in *Cell*, they have achieved just that in mice. In doing so, they reveal clues to the mechanism by which the hormone leptin promotes fat loss in mammals.

One of the main functions of one type of fat, white adipose tissue (WAT), is to store lipids. WAT is also the primary source of leptin, which is secreted in response to lipid storage and acts in the brain to reduce body-fat mass[2,3]. Although many experiments[4] have shown that leptin activates lipolysis (lipid breakdown), the mechanisms that underlie this feedback loop are less well defined. In particular, although lipolysis is thought to be under tight control of the brain and the peripheral nervous system[4], several key questions remain unanswered. For example, does WAT receive bona fide innervation from the autonomic nervous system (the part of the peripheral nervous system that regulates day-to-day organ function)? And how are fat depots slimmed down when the brain is instructed that fat stores are more than sufficient?

Zeng and colleagues used state-of-the-art techniques to investigate whether the lipolytic effect of leptin is mediated by the autonomic nervous system. Technical innovations[5,6] now allow researchers to clear intact organs of lipids, making the organs more transparent and thus amenable to visualization by microscopy. The authors exploited this advance to clear mouse-derived inguinal fat pads (masses of closely packed fat cells close to the hind leg), and then used sophisticated imaging techniques to reconstruct 3D anatomical pictures of the entire tissue[7]. This reconstruction revealed that thick bundles of neuronal projections called axons cover the surface of the fat pad.

The researchers found that these bundles belong to the sympathetic nervous system — the part of the autonomic nervous system that stimulates the fight-or-flight response, and that is responsible for accelerating heart rate, dilating pupils and activating sweat secretion. Indeed, many of the bundles expressed the enzyme tyrosine hydroxylase, which helps to synthesize catecholamine molecules such as noradrenaline that act as neurotransmitters in the sympathetic nervous system. Zeng and colleagues also showed *in vivo* that fat cells were located close to nerve fibres that expressed tyrosine hydroxylase. Fat pads were not analysed using electron microscopy, which could have verified whether sympathetic neurons terminate on fat-cell membranes. But these data nonetheless indicate that tyrosine-hydroxylase-expressing axonal projections make contact with some fat cells.

Next, Zeng *et al.* investigated the relationship between activation of the axon bundles and fat-cell metabolism using optogenetics — a revolutionary technique in which light-sensitive ion-channel proteins are selectively expressed in certain neurons and activate those neurons when exposed to light[8]. Although the technique is commonly used on the brain, optogenetic experiments on other tissues are often hindered by the fact that neurons outside the brain can have long axons; this means that high levels of light-controlled ion-channel-protein expression are required to drive photoactivation of the distant projections[9]. Exacerbating this problem, the axons that innervate the inguinal fat pad originate in clusters of neuronal cell bodies that are almost impossible to access for precise, chronic light stimulation.

The authors overcame these technical challenges by using genetic techniques to specifically target sympathetic axons, and locally modulated the activity of axons innervating the fat pad. As a compelling verification of the method's effectiveness, illuminating the inguinal fat had the same effect as treating mice with leptin — levels of noradrenaline increased, as did phosphorylation (an activating molecular modification) of hormone-sensitive lipase (HSL), an enzyme that the authors used as a measure of leptin-elicited lipolysis. Daily optogenetic activation of axons over several weeks reduced fat mass. Conversely, disrupting neuronal input to the fat pad genetically, surgically or pharmacologically almost completely blocked leptin-evoked HSL phosphorylation. This indicates clearly that leptin-triggered lipolysis depends on activation of the sympathetic neurons that project to fat (Fig. 1).

To investigate the molecular mechanism underlying this response to leptin, Zeng *et al.* analysed genetically engineered mice in which catecholamine signalling was blocked. The mice lacked either an enzyme involved in noradrenaline synthesis or isoforms of noradrenaline-receptor proteins called β-adrenoceptors, which are expressed by fat cells. Although leptin treatment resulted in phosphorylated HSL and fat loss in wild-type mice, these effects were attenuated in both types of mutant. Notably, mice lacking the β-adrenoceptor isoforms β1 and β2 showed more lipase phosphorylation and whole-body fat loss than those lacking β1, β2 and β3, consistent with a study[10] that pointed to a dominant role for β3 receptors in lipolysis.

Finding that sympathetic neurons innervate

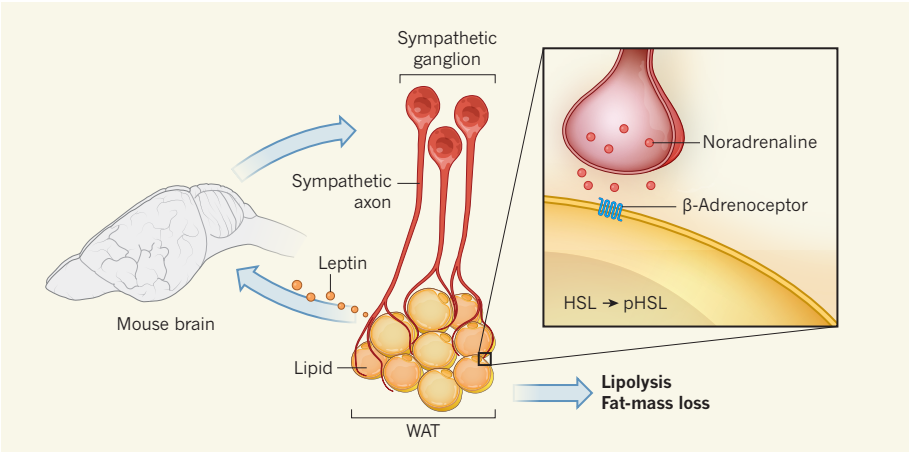


Figure 1 | Sympathetic to fat loss. The hormone leptin is secreted from fat tissue called white adipose tissue (WAT) in response to lipid storage. Zeng *et al.*[1] report that, in mice, leptin acts in the brain, triggering signals that activate ganglionic neurons of the sympathetic nervous system whose projections (called axons) wrap around fat cells. The neurons release the neurotransmitter molecule noradrenaline, which signals to β-adrenoceptor proteins on the fat cells. This promotes phosphorylation (p) of the enzyme hormone-sensitive lipase (HSL), triggering lipolysis (lipid breakdown) and so fat loss.

WAT and mediate leptin-stimulated lipolysis is not surprising. However, Zeng and colleagues' study fills a gap in our understanding of precisely how organisms respond to an abundance of leptin. Their work also specifically demonstrates that sympathetic neurons projecting to WAT are a central trigger for leptin-mediated lipolysis.

Of course, questions arise from these findings. Leptin is thought to signal through several brain areas[11], but it remains unclear which neuronal networks sense increased blood leptin concentrations and control sympathetic relay stations to ultimately regulate lipolysis and fat mass. Notably, only half of the nerve fibres found in WAT expressed tyrosine hydroxylase, and the authors did not analyse the other half, nor the characteristics of the fat cells that the neurons innervate. Although their identities remain elusive, these neurons and fat cells hold the potential for further exciting discoveries. Future experiments should define the key brain areas that control sympathetic traffic to WAT and the molecular circuitry that controls lipolysis downstream of these effectors.

Zeng *et al.* estimated that tyrosine-hydroxylase-expressing neurons envelop between 3 and 12% of fat cells, a relatively sparse coverage. Nonetheless, the fact that optogenetic activation markedly increased lipolysis indicates that catecholamine signalling through neuro-adipose junctions has an important role in the control of lipid homeostasis. Given that leptin resistance is a common feature of obesity, it is to be hoped that this study will fuel further dissections of the brain–fat axis. It might also open a door to assessing the therapeutic potential of controlling catecholamine signalling in fat. ■

Johan Ruud *and* **Jens C. Brüning** *are at the Max Planck Institute for Metabolism Research, Cologne 50931, Germany.*
e-mails: johan.ruud@sf.mpg.de; bruening@sf.mpg.de

1. Zeng, W. *et al. Cell* **163,** 84–94 (2015).
2. Zhang, Y. *et al. Nature* **372,** 425–432 (1994).
3. Friedman, J. M. & Halaas, J. L. *Nature* **395,** 763–770 (1998).
4. Bartness, T. J., Liu, Y., Shrestha, Y. B. & Ryu, V. *Front. Neuroendocrinol.* **35,** 473–493 (2014).
5. Chung, K. *et al. Nature* **497,** 332–337 (2013).
6. Ke, M.-T., Fujimoto, S. & Imai, T. *Nature Neurosci.* **16,** 1154–1161 (2013).
7. Sharpe, J. *et al. Science* **296,** 541–545 (2002).
8. Sohal, V. S., Zhang, F., Yizhar, O. & Deisseroth, K. *Nature* **459,** 698–702 (2009).
9. Atasoy, D., Aponte, Y., Su, H. H. & Sternson. S. M. *J. Neurosci.* **28,** 7025–7030 (2008).
10. Gettys, T. W., Harkness, P. J. & Watson, P. M. *Endocrinology* **137,** 4054–4057 (1996).
11. Myers, M. G. Jr & Olson, D. P. *Nature* **491,** 357–363 (2012).

MICROBIOLOGY

Electrical signalling goes bacterial

The discovery that potassium ion channels are involved in electrical signalling between bacterial cells may help to unravel the role of ion channels in microbial physiology and communication. SEE ARTICLE P.59

SARAH D. BEAGLE & STEVE W. LOCKLESS

Biological membranes separate cells or cellular compartments from the rest of the world, protecting the internal contents from the sometimes hostile, and always different, external milieu. However, cells are not closed systems and must pass information and matter, including ions, selectively across the membrane barrier. Proteins called ion channels facilitate the movement of ions across the membrane by allowing each ion to flow passively down its electrochemical gradient. Although ion channels mediate rapid, long-range communication in eukaryotes (the group of organisms that includes plants, animals and fungi), a signalling role for bacterial ion channels has remained elusive[1,2]. In this issue, Prindle *et al.*[3] (page 59) report the first example of a bacterial potassium channel that functions in a signalling role, through long-range coordination of metabolic oscillations.

The current study is an extension of the same laboratory's previous discovery[4] that adherent communities of *Bacillus subtilis* bacteria, known as biofilms, grow in periodic cycles once the colony reaches a threshold size (Fig. 1). The authors proposed that these oscillations arise when the cells in the biofilm's interior become deprived of glutamate, owing to high consumption of the amino acid by peripheral cells. Glutamate starvation in the interior cells reduces their production of ammonium ions, which the peripheral cells need, resulting in arrested cell growth in the periphery. Following replenishment of glutamate in the interior cells, ammonium production increases, leading to growth of peripheral cells. The linked metabolic processes of cells within the biofilm community raised the question of how the metabolic state of cells is communicated over long distances.

Maintenance of the proper intracellular concentrations of glutamate and ammonium depends on the electrical potential across the cell membrane[5,6], known as the membrane potential. Therefore, Prindle *et al.* investigated whether electrical signalling is responsible for the long-range coordination of metabolic oscillations across the bacterial population. Using a voltage-sensitive fluorescent dye, the authors detected rhythmic synchronized fluctuations in membrane potential across the biofilm. Eliminating the need for glutamate and ammonium by adding the amino acid glutamine to the cells' growth medium quenched these fluctuations, thereby linking electrical signalling and metabolism.

The observed changes in the membrane potential could result from sodium ions (Na^+) moving into the cell, potassium ions (K^+) moving out, or a combination of both. Using fluorescent dyes that specifically bind to either Na^+ or K^+, the researchers found a direct correlation between the timing of K^+ efflux and changes in membrane potential, suggesting that K^+ efflux might propagate signals across the biofilm.

Because the K^+ channel YugO is involved in *B. subtilis* biofilm formation[7], Prindle *et al.* next asked whether this channel mediates K^+ efflux. As expected, glutamate limitation in wild-type cells led to K^+ efflux, whereas no K^+ efflux was observed in cells lacking the *yugO* gene. Similarly, deletion of the TrkA domain of YugO, which gates K^+ flux, decreased the propagation of electrical oscillations under limited glutamate conditions. These results indicate that YugO is activated by glutamate limitation and is required to propagate the K^+ signal through the biofilm (Fig. 1). The use of extracellular K^+ to propagate a metabolic stress signal through the bacterial community is reminiscent of the increase in extracellular K^+ that drives the dilation of blood vessels in the mammalian brain[8] in response to stress, suggesting that some K^+ channels in bacteria and eukaryotes have evolved to accomplish similar outcomes.

Prindle and colleagues' study establishes the first example of a signalling function for a bacterial K^+ channel. Although previous studies[9,10] have established a role for various classes of bacterial channel in regulating cellular osmotic pressure, the impressive evolutionary conservation of eukaryotic and bacterial channels at the protein-sequence and structural levels provides additional evidence that some bacterial ion channels probably have signalling roles[11–13]. It is notable that the first demonstration of a signalling role for bacterial ion channels occurs in the context of bacteria acting as multicellular entities, and that this function serves to coordinate the

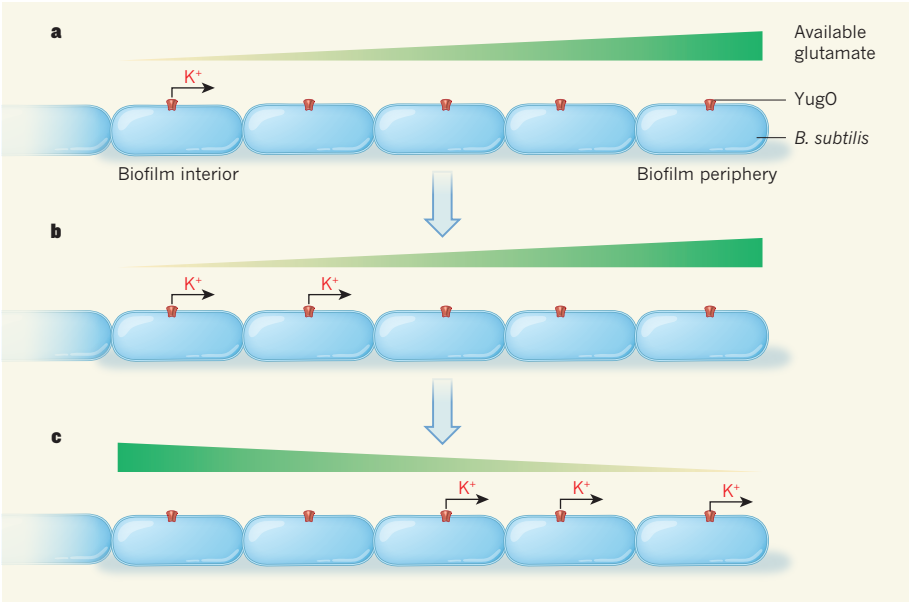


Figure 1 | Shocking communication. *Bacillus subtilis* bacteria can form communities called biofilms, in which cells both in the interior and on the periphery require the amino acid glutamate to survive and grow. **a**, When peripheral cells take up most of the available glutamate, the interior cells become starved. Prindle *et al.*[3] propose that nutrient-stressed interior cells secrete potassium ions (K^+) through the YugO K^+ channel. **b**, The release of K^+ ions then changes the transmembrane voltage of cells and leads to the subsequent release of K^+ ions from neighbouring cells, propagating the starvation signal. **c**, The signal propagation ultimately reduces the uptake of glutamate in peripheral cells. Glutamate becomes available for interior cells to consume and the cycle is reset.

metabolic states of neighbouring cells.

Unlike a eukaryotic action potential (in which electrical signal propagation is fast, owing to the rapid rising and falling of the membrane potential), the signalling that coordinates metabolic oscillations in *B. subtilis* occurs over a longer time period. Using mathematical modelling, the authors provide evidence that K^+ efflux alone can account for the slow signal propagation. This propagation, which is perhaps an evolutionary precursor to the faster action potential, seems to retain overall biofilm stability by synchronizing the growth of peripheral cells and metabolic maintenance of interior cells.

Although the present study highlights similar signalling roles for eukaryotic and bacterial ion channels, many questions remain to be addressed. What are the metabolic intermediates that activate YugO following glutamate starvation in *B. subtilis*? One possibility is that the TrkA regulatory domain senses the energy level of the cell by binding ATP or ADP, two molecules that have been shown[14] to regulate the TrkA protein in other bacteria. What is the magnitude of the changes in membrane potential, and how does this affect other voltage-dependent processes in the membrane? More generally, it will be interesting to determine whether this mechanism is used by other community-forming species as a way to regulate metabolism and growth.

The discovery of this K^+ signalling mechanism highlights the complexities of bacterial social communication. Most bacterial cells are too small to require electrically propagated intracellular signalling; instead, diffusion of signalling molecules in the cytoplasm is sufficiently rapid. It remains to be seen whether other bacterial social behaviours are governed by electrical signalling. Perhaps such signalling plays a part in interspecies communication, for instance between biofilms and epithelial cells in the gut. Like predator–prey interactions in some of the more complex eukaryotic species, it could be that microorganisms compete with each other by secreting toxins that interfere with an adversary's ion-channel activities. Finally, the fact that signalling in the biofilm shares several characteristics with electrical signalling in the nervous system — including the use of the common neurotransmitter molecule glutamate — highlights an exciting functional connection between these evolutionarily distant systems. ■

Sarah D. Beagle *and* **Steve W. Lockless** *are in the Department of Biology, Texas A&M University, College Station, Texas 77843-3474, USA.*
e-mail: lockless@bio.tamu.edu

1. Kubalski, A. & Martinac, B. (eds) *Bacterial Ion Channels and Their Eukaryotic Homologs* (ASM Press, 2005).
2. Booth, I. R., Edwards, M. D. & Miller, S. *Biochemistry* **42,** 10045–10053 (2003).
3. Prindle, A. *et al. Nature* **527,** 59–63 (2015).
4. Liu, J. *et al. Nature* **523,** 550–554 (2015).
5. Boogerd, F. C. *et al. FEBS Lett.* **585**, 23-28 (2011).
6. Tolner, B., Ubbink-Kok, T., Poolman, B. & Konings, W. N. *J. Bacteriol.* **177,** 2863–2869 (1995).
7. Lundberg, M. E., Becker, E. C. & Choe, S. *PLoS ONE* **8,** e60993 (2013).
8. Filosa, J. A. *et al. Nature Neurosci.* **9,** 1397–1403 (2006).
9. Levina, N. *et al. EMBO J.* **18,** 1730–1737(1999).
10. Epstein, W. *Prog. Nucleic Acid Res.* **75,** 293–320 (2003).
11. Iyer, R., Iverson, T. M., Accardi, A. & Miller, C. *Nature* **419,** 715–718 (2002).
12. MacKinnon, R., Cohen, S. L., Kuo, A., Lee, A. & Chait, B. T. *Science* **280,** 106–109 (1998).
13. Chen, G. Q., Cui, C., Mayer, M. L. & Gouaux, E. *Nature* **402,** 817–821 (1999).
14. Cao, Y. *et al. Nature* **496,** 317–322 (2013).

This article was published online on 21 October 2015.

QUANTUM PHYSICS

Quantum sound waves stick together

A sensitive cold-ion experiment probes sound at the level of phonons, the fundamental quantum units of vibration. It shows that phonons mix in such a way that they can be classified as 'bosonic' particles, like photons. SEE LETTER P.74

DAVE KIELPINSKI

The phenomenon of wave interference is observed in various settings, including optics, electronics and acoustics. In constructive interference, the crests and troughs of interfering waves reinforce each other, whereas in destructive interference they cancel each other out. Although we think of sound as consisting of macroscopic waves, it has a quantum nature. The energy of a sound wave is an integer multiple of a fundamental quantum of vibrational energy called a phonon. On page 74 of this issue, Toyoda *et al.*[1] report the effect of two-phonon interference, and show that the interfering phonons 'stick together' — they are never observed to go different ways.

The interference of sound waves is not just of academic interest. For instance, it is the operating principle of noise-cancelling

headphones. These create their own sound vibrations, which are tuned to destructively interfere with external vibrations. The two vibrations cancel at the ear, and so no sound is heard. By contrast, they constructively interfere at other locations, away from the ear.

On a smaller scale, a quantum-mechanics principle dictates that when the number of phonons (or particles such as photons or electrons) is accurately known for a system, the locations of the crests and troughs of the waves associated with these particles cannot be known with certainty. However, one can still perform an interference experiment to see what happens.

In 1987, the physicists Hong, Ou and Mandel demonstrated[2] that, surprisingly, the interference of two photons is either completely constructive or completely destructive, and that the two possibilities coexist until the result of the experiment is observed. Two detectors that register photons at two output ports always measure zero photons at one port and two photons at the other. This is known as the Hong–Ou–Mandel effect. By contrast, in the case of electrons, it has been shown[3] that two electrons will never register at the same port — they always go their separate ways.

Toyoda and colleagues use highly sophisticated experimental-physics techniques to probe sound at the level of individual phonons. At room temperature, the atoms in matter show random thermally driven vibrations that act as background 'noise', overwhelming the quantum effects of sound. Only matter that has been cooled to near absolute zero temperature displays sufficiently small thermal vibrations to allow such effects to be measured.

To perform these measurements, the authors use two calcium ions that have been electromagnetically trapped in a chamber under ultrahigh-vacuum conditions. Heat cannot reach the ions because they are not in contact with the chamber's walls and there is no gas in the chamber to transfer it. Toyoda *et al.* suppress the ions' residual thermal vibrations using a technique known as laser cooling, allowing the ions' quantized vibrations (the phonons) to be revealed. By applying appropriately tuned laser pulses to the ions, they can then add or remove vibrations, one phonon at a time. A follow-up sequence of pulses causes the ions to fluoresce only if they are vibrating, and the authors use the detected fluorescence as an optical marker for sound at the quantum level.

To observe the interference of two phonons, Toyoda *et al.* start by 'feeding' one phonon to each ion. Because the two ions are positively charged, they repel each other, so that when one vibrates, the other one gently wiggles. This wiggling effect causes a phonon that starts out on its own ion to mix slowly with the other ion's phonon, and so to interfere with it. The authors observe that, almost always, both phonons end up on the same atom, but which one? The phonons don't care. In this situation, quantum mechanics predicts that both phonons reside together on one atom, and at the same time, both reside on the other atom — at least, as long as no one measures the location of the phonons. By adding an extra interference step to the experiment, the authors obtain substantial evidence that the phonons can, in fact, seem to be in both places at once.

The effects reported by the authors might be used in the quantum engineering of acoustic devices.

The authors' results are a crucial test of the quantum theory of sound, and definitively prove that phonons are bosons rather than fermions (bosons, such as photons, are particles that have integer spin angular momentum, whereas fermions, such as electrons, have half-integer spin). Every quantum system falls into one of these two categories, and this classification has physical ramifications. For instance, laser-like wave emission commonly occurs in systems that have a large number of bosons — it has been observed for phonons in trapped-ion experiments[4] and in microscopic devices known as toroidal resonators[5].

In optical systems, the Hong–Ou–Mandel effect powers applications such as quantum computing, simulation and sensing[6]. The current work indicates that phononic systems could also be suitable for quantum-enhanced applications. Nanometre-scale mechanical systems, although more vulnerable to thermal noise than trapped ions, offer a wider field of potential quantum phononics applications because they can operate at room temperature and under atmospheric pressure. Recently[7], it has become possible to control and measure single phonons in nanomechanical systems — interference experiments may soon follow. As these systems become increasingly complex, the effects reported by Toyoda *et al.* might be used in the quantum engineering of acoustic devices and circuits. ■

Dave Kielpinski *is at Hewlett Packard Laboratories, Palo Alto, California 94304, USA.*
e-mail: david.kielpinski@hpe.com

1. Toyoda, K., Hiji, R., Noguchi, A. & Urabe, S. *Nature* **527,** 74–77 (2015).
2. Hong, C. K., Ou, Z. Y. & Mandel, L. *Phys. Rev. Lett.* **59,** 2044–2046 (1987).
3. Bocquillon, E. *et al. Science* **339,** 1054–1057 (2013).
4. Vahala, K. *et al. Nature Phys.* **5,** 682–686 (2009).
5. Grudinin, I. S., Lee, H., Painter, O. & Vahala, K. J. *Phys. Rev. Lett.* **104,** 083901 (2010).
6. O'Brien, J. L., Furusawa, A. & Vučković, J. *Nature Photon.* **3,** 687–695 (2009).
7. Cohen, J. D. *et al. Nature* **520,** 522–525 (2015).

GENE REGULATION

Expression feels two pulses

Single-cell analyses reveal that combinatorial changes in the intracellular locations of transcription factors can tune the expression of the factors' target genes in response to environmental stimuli. SEE ARTICLE P.54

ANTOINE BAUDRIMONT & ATTILA BECSKEI

Most transcription factors exert their action continuously, but some act in pulses by moving rapidly in and out of the nucleus. Transmitting cellular signalling-pathway information in pulses or oscillations has several advantages over continuous signalling. For example, information can be encoded in the frequency or amplitude of pulsing, boosting the amount of information transmitted. Investigating this phenomenon has proved difficult, however, because the behaviour of pulsatile transcription factors varies greatly from cell to cell[1]. In this issue, Lin *et al.*[2] (page 54) overcome this hurdle and demonstrate that combinations of transcription-factor pulses that change in response to environmental stimuli can regulate gene expression.

The Msn2 protein, which is expressed in the budding yeast *Saccharomyces cerevisiae*, was the first transcription factor to be identified as pulsatile, moving to the nucleus to activate transcription when cells are exposed to light[3]. Although such pulsatile signalling patterns can be advantageous, they are also prone to disruption, because the message transmitted varies as time passes. In electronics, such problems are typically solved by ensuring that more than one component can perform the same task. Theoretically, the same principles apply to cell signalling — propagating signalling pulses through multiple pathways that are then reintegrated is predicted to improve reliability[4]. Indeed, Msn2 is known[5] to act with the pulsatile transcriptional repressor protein Mig1 to control gene expression in response to various stresses.

Lin *et al.* analysed the dynamics of Msn2

and Mig1 pulses by generating strains of *S. cerevisiae* in which the two transcription factors were tagged by different fluorescent proteins, allowing their intracellular locations to be tracked. The authors attached these cells to a microfluidic device through which cell-growth media were passed, and monitored transcription-factor movements as well as any subsequent changes in the transcription of genes whose expression is regulated by both factors.

Depleting glucose in the cell media triggered the export of Mig1 from the nucleus and the import of Msn2, increasing the expression of target genes. If Msn2 and Mig1 acted according to a simple continuous regulatory scheme, or if they were pulsatile but the timing of pulses was completely random, then glucose depletion would gradually alter the average level of each transcription factor in the nucleus across the population of cells, and the expression of target genes would gradually increase to a new steady-state level. Instead, however, the authors observed a 'transient phase' immediately after glucose depletion, during which the average nuclear levels of Msn2 and Mig1 were higher and lower, respectively, than when they subsequently reached steady-state levels (Fig. 1a). An overshoot such as this is often observed when systems adapt to change[6], and it can decrease the time it takes for target-gene expression levels to reach the new steady state. Indeed, a kinetically similar response is known to occur[7] when glucose concentration increases: target-gene expression is repressed by Mig1, lowering levels of the corresponding RNA transcript, but a transient destabilization of the transcript helps to speed up the process by promoting transcript degradation.

Lin and colleagues hypothesized that the overshoot they observed was not just a transient event, but might persist in steady-state conditions in the form of pulses that could not be observed on a population-wide level because their effects averaged out. When analysing single cells, however, the authors found that the pulsing of each transcription factor was sporadic and irregular under steady-state environmental conditions, making it hard to define individual pulses. In principle, many criteria could be used to define such pulses, but most would be of little practical relevance. The authors developed an interesting and pragmatic approach to detecting individual pulses, based on a neuroscience technique called spike-triggered averaging[8]. In their adapted version, which the authors dubbed pulse-triggered averaging, pulses were measured as averages that were based on the dynamics of Mig1 and Msn2 over a set time period around peaks in nuclear Msn2 levels.

Justifying their approach, the authors demonstrated that a target gene responded to changes in Msn2 or Mig1 that the technique registered as pulses. For instance, Msn2 pulses were followed by elevated gene expression if there was no overlap with a Mig1 pulse (that is, if Mig1 was not in the nucleus at the same time). Conversely, there was a decrease in target-gene expression if the Msn2 pulse was counteracted by an overlapping Mig1 pulse. These observations confirm that the transcription-factor pulses do indeed persist under steady-state conditions (Fig. 1b).

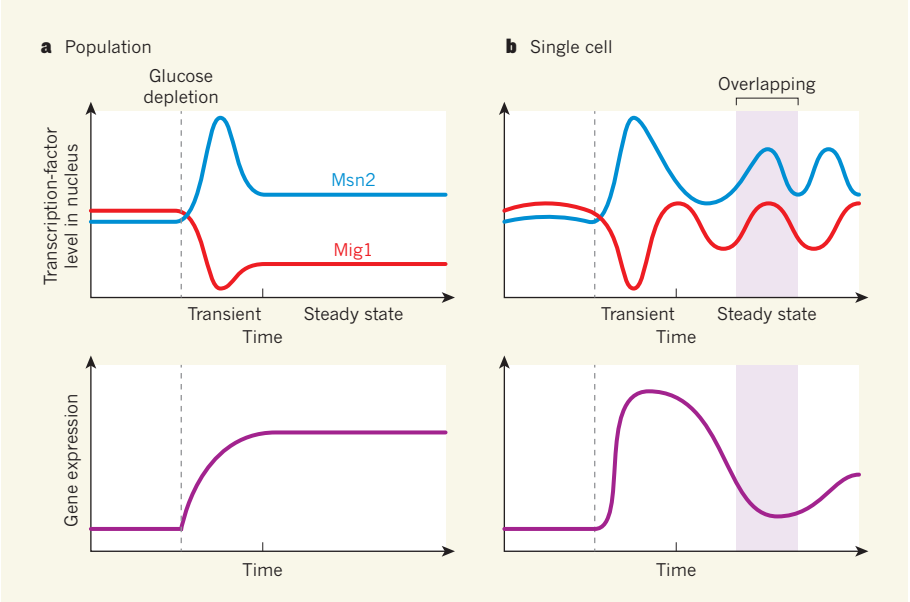


Figure 1 | Interpreting transcription-factor pulses. The transcription factors Msn2 and Mig1 enter the nucleus in pulses to respectively activate and repress transcription of the same target genes. **a**, Lin *et al.*[2] report changes in transcription-factor pulsing in response to environmental stimuli. A decrease in glucose concentration in the medium around the cell causes a large transient decrease in the level of Mig1 in the nucleus, and a rapid increase in Msn2. This 'overshoot' allows cells to adapt to change by promptly increasing gene expression. After this transient phase, nuclear levels of each factor, and hence gene expression, remain steady on a population-wide level. **b**, The authors show that it is only in the transient phase that all cells display synchronized non-overlapping pulses. After this phase, levels of each factor pulse randomly in single cells. When pulses of both factors overlap in the nucleus, gene expression falls. However, when pulses of Msn2 do not overlap with Mig1, expression increases.

At elevated steady-state glucose concentrations, the percentage of overlapping pulses increased to a level beyond that expected by chance, enhancing the efficiency with which target-gene expression was repressed. By contrast, during the transient phase, Lin and colleagues detected only non-overlapping pulses, suggesting that during this period the timing of pulses is modulated. In this way, pulses are synchronized between cells, resulting in an overshoot at the population-wide level.

Pulse-triggered averaging could now become a powerful tool for analysing other regular oscillating reactions in cells. So far, most studies have focused on average cell behaviour, but cell-cycle checkpoints, for instance, elicit single-cell responses with considerable cell-to-cell variability[9]. Pulse-triggered averaging may help to disentangle the underlying regulatory interactions.

Moreover, the authors' approach makes it possible to analyse gene regulation without understanding all of a system's parameters. For instance, the current study showed not only that a fully overlapping repressor pulse can neutralize an activating pulse, but also that the same neutralization can occur when the two pulses are separated by a few minutes, without needing to understand the root causes. It is important to note that not all Msn2 pulses elicited target-gene expression, even when Mig1 activity was low. This provides a reminder that mass-action kinetics, stochastic modelling and identification of reaction mechanisms must be included in complete models of gene regulation. Research into these topics has undergone marked development in recent years, and may soon converge, making it possible to understand the dynamics of signalling pathways in detail. ■

Antoine Baudrimont *and* **Attila Becskei**
are in the Biozentrum, University of Basel, CH-4056 Basel, Switzerland.
e-mail: attila.becskei@unibas.ch

1. Corrigan, A. M. & Chubb, J. R. *Curr. Biol.* **24,** 205–211 (2014).
2. Lin, Y., Sohn, C. H., Dalal, C. K., Cai, L. & Elowitz, M. B. *Nature* **527,** 54–58 (2015).
3. Jacquet, M., Renault, G., Lallet, S., De Mey, J. & Goldbeter, A. *J. Cell Biol.* **161,** 497–505 (2003).
4. Hansen, A. S. & O'Shea, E. K. *eLife* **4,** e06559 (2015).
5. De Wever, V., Reiter, W., Ballarini, A., Ammerer, G. & Brocard, C. *EMBO J.* **24,** 4115–4123 (2005).
6. Drengstig, T., Ueda, H. R. & Ruoff, P. *J. Phys. Chem. B* **112,** 16752–16758 (2008).
7. Hsu, C. *et al. Nature Commun.* **3,** 682 (2012).
8. Gerstner, W. *Neural Netw.* **14,** 599–610 (2001).
9. Liang, H. *et al. Nature Commun.* **5,** 4048 (2014).

This article was published online on 14 October 2015.

ARTICLE

doi:10.1038/nature16065

Australia is 'free to choose' economic growth and falling environmental pressures

Steve Hatfield-Dodds[1], Heinz Schandl[1], Philip D. Adams[2], Timothy M. Baynes[3], Thomas S. Brinsmead[4], Brett A. Bryan[5], Francis H. S. Chiew[1], Paul W. Graham[4], Mike Grundy[6], Tom Harwood[1], Rebecca McCallum[1], Rod McCrea[7], Lisa E. McKellar[7], David Newth[8], Martin Nolan[5], Ian Prosser[1]† & Alex Wonhas[3]

Over two centuries of economic growth have put undeniable pressure on the ecological systems that underpin human well-being. While it is agreed that these pressures are increasing, views divide on how they may be alleviated. Some suggest technological advances will automatically keep us from transgressing key environmental thresholds; others that policy reform can reconcile economic and ecological goals; while a third school argues that only a fundamental shift in societal values can keep human demands within the Earth's ecological limits. Here we use novel integrated analysis of the energy–water–food nexus, rural land use (including biodiversity), material flows and climate change to explore whether mounting ecological pressures in Australia can be reversed, while the population grows and living standards improve. We show that, in the right circumstances, economic and environmental outcomes can be decoupled. Although economic growth is strong across all scenarios, environmental performance varies widely: pressures are projected to more than double, stabilize or fall markedly by 2050. However, we find no evidence that decoupling will occur automatically. Nor do we find that a shift in societal values is required. Rather, extensions of current policies that mobilize technology and incentivize reduced pressure account for the majority of differences in environmental performance. Our results show that Australia can make great progress towards sustainable prosperity, if it chooses to do so.

Our analysis uses a new integrated multi-model framework developed for the Australian National Outlook[1]. Australia is globally relevant: a major exporter of energy, mineral and agricultural products, with high per capita income, greenhouse gas emissions, water extractions, and habitat loss. The framework assesses energy–water–food interactions (and links to ecosystem services) in the context of climate change[2], and uses more than 20 scenarios to explore a diverse range of factors shaping future Australian economic and environmental outcomes[1,2]. Interacting national trends and policies include energy and resource efficiency, agricultural productivity, consumption and working hours, and new land-sector markets for energy feed-stocks and ecosystem services (carbon sequestration and biodiversity conservation). These are modelled against four levels of national and global greenhouse gas emissions reduction effort (from no abatement to very strong abatement), and associated global climate trajectories (see Extended Data Fig. 9). As well as assessing the range of scenario outcomes, we identify the relative contributions of different types of choices. 'Collective choices' are defined as decisions that can only be implemented by groups of actors, and then constrain or empower 'individual choices' (particularly through changing rules and institutions). For example, individual choices about whether to drive or catch a train to work are strongly shaped by prior collective choices about transport infrastructure.

The framework accounts for detailed interactions across sectors and spatial scales. The focal scale is national (the continent of Australia), accounting for key processes at higher (global) and lower (sub-national) spatial scales. This cross-domain integrated approach is needed because partial assessments may not account for constraints or adverse impacts that would undermine an otherwise 'sustainable' trajectory[3–8]. The projections and indicators are fully consistent with the international System of National Accounts[9]. We provide more details in the Supplementary Methods (section 'Overview of modelling framework and scenarios') and results for more than 60 national and global indicators in the Supplementary Data.

Novel aspects of the analysis include assessing the potential for markets for ecosystem services to supply carbon sequestration and habitat restoration (and implications for agricultural output[7] and extinction risk)[10,11]; assessing future water stress rather than simple volume of water extracted[2,12]; exploring material extractions and environmental footprints[13]; and integrating these elements with established models for analysing energy, greenhouse gas emissions and economic performance[2,14–17]. We are not aware of any other future-looking modelling that integrates this range of issues and indicators (Supplementary Methods, 'Overview of modelling framework and scenarios').

Economic and physical decoupling is possible

We find that substantial economic and physical decoupling is possible[18]. Economically, Australia can achieve strong economic growth to 2050, indicated by rising gross domestic product (GDP) and gross national income (GNI) per capita, in scenarios where environmental pressures fall or are stable. Physically, we find the services derived from natural resources (energy (Extended Data Fig. 2), water (Extended Data Fig. 3), food (Extended Data Fig. 4)) can increase, while associated environmental pressures ease (greenhouse emissions (Extended Data Fig. 6), water stress (Extended Data Fig. 3), native habitat loss (Extended Data Fig. 5)). Importantly, these projected decouplings do not involve a reduction in the value of Australia's heavy industry (Extended Data

[1]CSIRO, Black Mountain Laboratories, Acton, ACT 2601, Australia. [2]Victoria University, Flinders Street, Melbourne, VIC 3000, Australia. [3]CSIRO, Julius Avenue, North Ryde, NSW 2113, Australia. [4]CSIRO, Energy Centre, Mayfield West, NSW 2304, Australia. [5]CSIRO, Waite Campus, Urrbrae, SA 5064, Australia. [6]CSIRO, Queensland Biosciences Precinct, St Lucia, QLD 4067, Australia. [7]CSIRO, Ecosciences Precinct, Dutton Park, QLD 4102, Australia. [8]CSIRO, Yarralumla Laboratories, Yarralumla, ACT 2601, Australia. †Present address: Bureau of Meteorology, Childers Street, Canberra, ACT 2600, Australia.

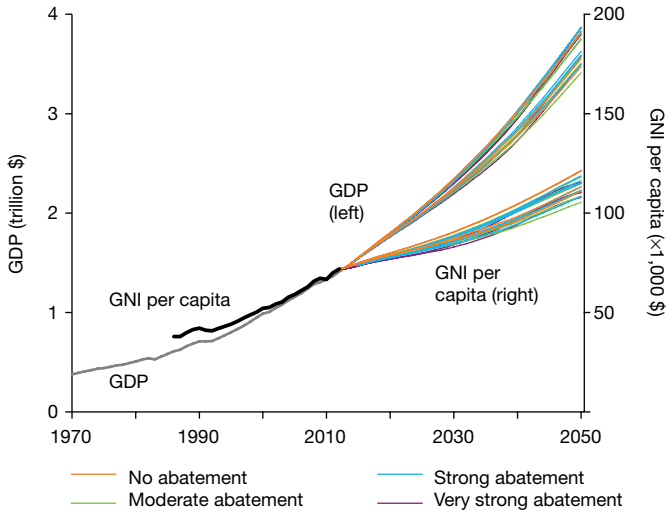


Figure 1 | Economic activity (GDP) and national income (GNI) continue to rise strongly in all scenarios. Projections for 20 scenarios. GDP measures the market value of goods and services produced. GNI here measures payments to national residents from domestic production (as foreign production is not modelled). All values are in real 2010 Australian dollars, adjusted for inflation; one trillion is defined as 1×10^{12}. Neither GDP or GNI is adjusted for changes in asset values, such as depreciation or the depletion of stocks of natural resources, and so do not measure pure income. More information on models and scenarios is provided in Supplementary Methods, 'Overview of modelling framework and scenarios'. Sources: Supplementary Data worksheets 1a and 1c.

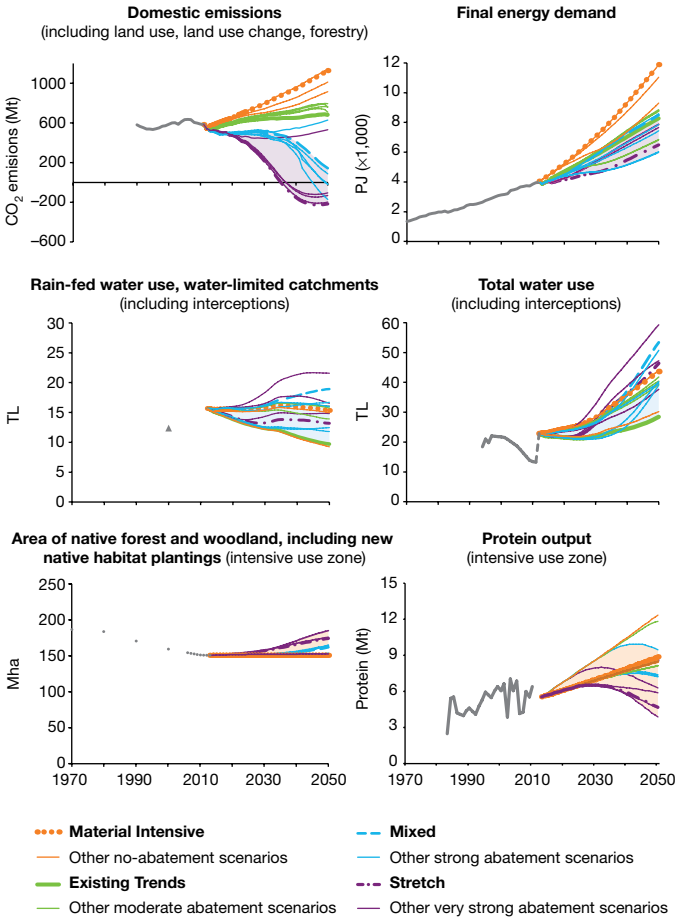


Figure 2 | Decoupling of emissions, water stress, and native habitat from the supply of energy, water and food, respectively, for 18–21 scenarios, 1970–2050. Each panel shows the scenario trajectories for a key indicator of resource use or environmental pressure. The shaded areas indicate scenarios in which environmental pressure decreases from current levels (in the left-hand panel), with the same scenarios shaded in the right hand panel of each row. Models and scenarios are described in Supplementary Methods, 'Overview of modelling framework and scenarios', and information on performance of multiple pressures across scenarios is provided in Supplementary Methods, 'Analysis of multiple pressures across scenarios'. Sources: Supplementary Data worksheets 6a, 2a, 3e, 3a, 5h and 4d.

Fig. 1g), or outsourcing its environmental footprint to other nations[13,19]. Instead energy- and material-intensive sectors are projected to increase their share of economic activity, even in scenarios with the strongest global abatement efforts[1,2].

In all scenarios, Australia's economy and living standards are projected to grow strongly (see Extended Data Fig. 1). As shown in Fig. 1, the value of economic activity (GDP) is projected to rise tenfold over the 80 years to 2050, driven by a 2.9-fold increase in population (Extended Data Fig. 8) and a 3.2–3.6-fold increase in GDP per capita (all values are in real 2010 Australian dollars, adjusted for inflation). National income (GNI) grows at a similar rate as GDP, with GNI per capita increasing by 58–82% from 2010 to 2050. Around two-thirds of the range of outcomes is explained by choices about working hours and consumption rather than environmental constraints. Average incomes rise by up to 66% if average working hours decline another 11% over the next four decades, in line with recent trends, and rise by 75% or more if there is no decline in working hours. The remaining income differential is accounted for by different assumptions and outcomes on resource efficiency, new land markets, agricultural productivity, and national and global abatement efforts.

Net greenhouse emissions show a clear decoupling from the growing economy, falling to zero or lower in some scenarios by 2040 (top row of Fig. 2). Australian emissions per capita could fall below the global average by 2050, from four times the global average today (Extended Data Figs 6b and 9f). One-third to one-half of Australia's projected emissions reductions are achieved through biosequestration from large areas of new carbon plantings (29–59 Mha in 2050, see Extended Data Fig. 5). The remainder is achieved by reducing the emissions- and resource-intensity of the economy. If there is a strong or very strong abatement effort, domestic emissions could fall by up to 33%, even as GDP grows more than 150%; and energy emissions could fall by up to 29% while energy use grows by 55–120%. Similarly, the total mass of fossil fuels, metals, non-metallic minerals and biomass[20] Australia uses is projected to decrease by 36% by 2050 in scenarios with very strong abatement and improved resource efficiency (Extended Data Fig. 1h). In other scenarios, total resource use is projected to increase by 69%[13].

National water extractions (by all sectors) are projected to increase by up to 101% by 2050. However, up to half (32–56%) of this water demand can be met by desalinisation in coastal cities and water recycling for industrial use. Water stress, indicated by rain-fed water use in water-limited catchments[12,21], improves or is stable in 7 of 18 scenarios (and is sensitive to governance of new carbon and biodiversity plantings, as noted below).

Pressures on biodiversity can also be reduced alongside economic growth and increased agricultural activity—resulting in increased native habitat and agricultural output volumes (including protein) in many scenarios[22] (bottom row of Fig. 2). Settings that give weight to biodiversity restoration could see mixed local native species plantings make up 36–47% of all carbon plantings in 2050 (against only 5% under a carbon-focused approach), increasing native habitat by up to 25% (37 Mha) in Australia's intensive use zone, and reversing the long-term trend. With strong abatement incentives, we find 11 Mha of habitat could be restored without large government outlays, reducing climate-related extinction risk by 7–9% (assessed for RCP 4.5 climate)[1].

However, these carbon and biodiversity plantings would reduce surface water flows, which could exacerbate pressures on river-based ecosystems in water-limited catchments (middle row of Fig. 2). Integrated

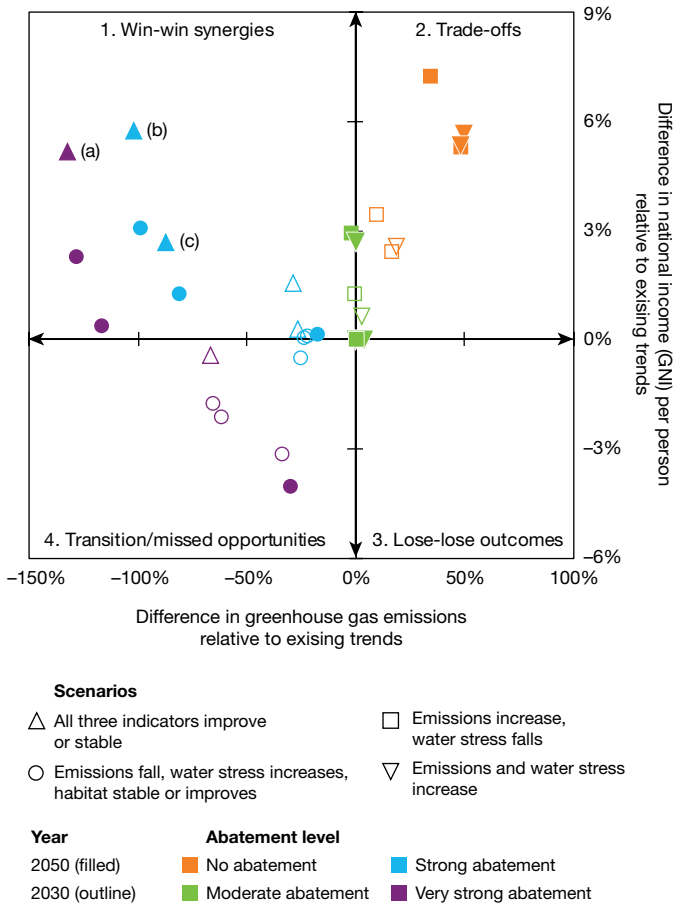


Figure 3 | Comparing living standards and emission outcomes across multiple scenarios. Differences in national income (GNI) and net greenhouse gas emissions in 2030 and 2050, relative to existing trends. Calculations based on 18 scenarios. Emissions, water stress and native habitat all improve or are stable in three scenarios, combining step change energy efficiency with very strong abatement (L1XI)—marked as (a)—or strong abatement (M3XI) (b), or trend energy efficiency with strong abatement (M3XR) (c). Differences shown are relative to existing trends (M2XR) controlling for working hours and consumption trends. Scenario assumptions and notation (such as M2XR) described in the text and Supplementary Methods, 'Calculations for Figure 3 and assessment of potential economic performance with different levels of global and national action to reduce greenhouse emissions'. Extended Data Fig. 6e shows time paths for each scenario from 2015 to 2050. Source: Supplementary Data worksheet 6e; see Extended Data Figs 1c and 6a.

governance is needed to properly balance their interceptions with competing extractive uses[23] (Supplementary Methods, 'Analysis of multiple pressures across scenarios'). Existing Australian governance arrangements cap extractions from water-limited catchments around current levels. The requirement to hold a water licence for new plantings embeds the price of water licences in these governance arrangements, as discussed below. Alternative governance assumptions could further restrain plantings, better safeguarding river health, but forgoing up to 0.5 Gt (5%) of cumulative national carbon sequestration by 2050.

Overall, two-thirds of the scenarios assessed (13 of 18) show improvement in at least one environmental indicator, but only three scenarios (all involving strong or very strong abatement and new land markets) show improvement or stable performance in all three environmental indicators, reflecting the tensions between reducing water stress and restoring terrestrial native habitat, and the importance of integrated governance (see Supplementary Methods Fig. 6 and Supplementary Methods, 'Analysis of multiple pressures across scenarios').

Policies to ease pressures extend established options

The scenario assumptions that result in reduced environmental pressures are all continuations of existing trends, combined with greater uptake of energy and water efficiency, and a shift towards stronger global and national greenhouse gas abatement (Supplementary Methods, 'Overview of modelling framework and scenarios'). Policy settings reflect market-based approaches that are already in place in Australia or other countries.

Greenhouse gas abatement is modelled as a uniform global broad-based carbon price, representing a variety of potential real-world mixes of regulation, standards, grants, taxes, or cap-and-trade arrangements. The carbon price in 2015 is US$15 (moderate scenario), US$30 (strong) and US$50 (very strong) per tonne of CO_2 emissions, and increases by around 4.5% per year in real terms (above inflation) to 2050. This drives a 90% reduction in the emissions intensity of Australian electricity from 2010 to 2050 in the stronger abatement scenarios (eliminating coal-fired electricity without carbon capture and storage before 2035 under the highest carbon price). Wholesale generation prices are 61–106% higher in 2050, and household electricity prices are 11–12% higher (strong) or 32% higher (very strong), compared to the no-abatement scenarios. However, affordability changes very little, owing to higher household incomes (in all scenarios) and higher energy efficiency in scenarios with higher prices[17].

Payments to Australian landholders for biosequestration are 15% below the global carbon price, with the forgone carbon revenue applied to increasing the share of native habitat plantings from 4–5% to 36–46% of total area in 2050. The resulting biodiversity 'top up payments' account for 22–30% of payments to habitat plantings in these scenarios over the decade to 2050, complementing carbon income. (These payments should be interpreted as a one-off payment for implementing a conservation covenant, for the area of new habitat added in that period.)

On water, we find that interceptions from new plantings result in increased water stress in many of the very strong abatement scenarios (which have the highest levels of new plantings). We find the profitability of carbon plantings is not sensitive to water licence prices: a doubling results in just a 4% reduction in the area of new plantings in water-limited catchments. Limiting the area of plantings to avoid this increased water stress would require a 200% increase in the water licence price (increasing the asset value of licences to existing owners).

Policy choices are crucial, not changes in values

These results provide insights into the contested relationship between economic growth and environmental sustainability[24], complementing historical analyses[18,25–27] (Supplementary Methods, 'Competing views on the prospects for sustainability'). A 'technological optimist' view considers market-driven technological advances will ensure that growth does not transgress key environmental thresholds[28–30]. Others suggest that institutional reform and new policies could achieve necessary changes within established values and paradigms[25,31–33], noting that environmental damage may occur during the long lags between problem identification and policy responses[18,25,34–36]. A third 'communitarian limits' view argues that sustainability will require a fundamental shift in societal values, often involving a rejection of economic growth[37,38], or a shift from consumerism to a values-based commitment to living within ecological limits[39].

We find that decoupling economic growth from environmental pressure before 2050 would not require a change in societal values, but is not automatic—contrary to both the communitarian limits and technological optimist positions. It is not projected to occur under existing trends, and requires, in our scenarios, collective choices to increase global and national abatement efforts.

The analysis explores potential behavioural change in several ways. The modelling simulates bottom-up individual choices on working hours and consumption that shape production and consumption as incomes rise (income elasticity) and relative prices change (price

elasticity). These choices interact with different assumptions about policy settings (reflecting collective choices), such as incentives for greenhouse gas abatement, and about bottom-up trends, such as the uptake of energy and water efficiency. None of the scenarios assume a new social or environmental ethic. In particular, increasing Australia's abatement effort in line with emissions reductions by other countries would be consistent with Australian public opinion[40] and assessments of Australia's national interest[41–43] in limiting the rise in average global temperature to 2 °C[5,7,32,44], and so is not interpreted as implying a change in values. Rather, the analysis reflects how goal-oriented human behaviour can change with circumstances (including new information, or changes in the actions of others), without requiring any change in underlying goals and values.

We find collective policy choices are crucial, explaining 46–94% of differences in environmental performance and resource use across the scenarios examined (see Extended Data Fig. 7 and Supplementary Methods, 'Assessing the contributions of individual and collective choices'). Consistent with the institutional reform approach[25,32,45,46], we find top-down collective choices are particularly important in shaping 'public good' outcomes—accounting for 83–94% of the differential in scenario outcomes for net greenhouse gas emissions, and 69–89% for greenhouse emissions excluding land sector sequestration. Bottom-up individual choices play a greater role when private and public benefits are aligned, such as when improved resource efficiency delivers financial savings. Individual choices account for up to half of the differential in scenario outcomes for energy use (33–47%) and non-agricultural water consumption (16–53%).

Giving value to natural assets can build new advantage

Economic analysis of climate change mitigation typically finds that limiting emissions involves near-term costs, but can yield net benefits over the long term (well after 2050) through avoided climate impacts[5,32,41,44]. Near-term co-benefits such as improved air quality and human health are also identified[47,48]. However, our analysis identifies additional near-term economic benefits for nations with a comparative advantage in ecosystem services, particularly carbon sequestration from reforestation. For these nations, stronger action to improve resource efficiency and environmental performance could unlock new sources of economic opportunity and growth, boosting near-term income while protecting natural assets essential to long-term well-being.

Figure 3 compares national income and net emissions outcomes in 2030 and 2050 for 18 scenarios. All seven stronger abatement scenarios (blue and purple) with land sector markets have better economic performance to 2050 than those with moderate abatement (green scenarios). National income (GNI) in 2050 in these scenarios is up to 6% higher than under existing trends (see quadrant 1). These win-win outcomes occur because carbon sequestration becomes more profitable than beef and other agricultural production across large areas of Australia (up to 58 Mha, or 70% of the intensive-use zone), in a world taking stronger action to reduce emissions. Stronger abatement incentives also promote electrification and the use of biofuels in road transport, reducing oil imports. These economic gains outweigh the costs of more stringent national emissions targets, as well as the impacts of lower global demand for (and value added from) Australia's emissions-intensive exports, relative to moderate national and global abatement (see Supplementary Methods, 'Calculations for Fig. 3 and assessment of potential economic performance with different levels of global and national action to reduce greenhouse emissions' and Extended Data Fig. 1i).

Across the scenarios explored, we find land-sector markets are needed to exploit these shifts in comparative advantage. Quadrant 4 reflects missed opportunities, including the scenario where very strong abatement action without land-sector markets leads to the worst relative economic performance (solid purple circle). Other scenarios in this quadrant involve transitions: pathways where emissions reductions generate net costs around 2030, but net benefits by 2050, relative to existing trends (see Extended Data Fig. 6e for time paths).

Quadrant 2 shows the scenarios in which there is no global or national action to reduce emissions, reflecting a decline from current modest abatement efforts. Here, national income in 2050 is projected to be 5–7% higher than for existing trends, while emissions are projected to be 35–51% higher. These scenarios illustrate the classic 'unsustainable development' trade-off, where higher near-term living standards are achieved at the cost of increased risks and future damage to the Earth's natural capital and life-support systems[5,46]. Adverse environmental feedbacks might see these scenarios shift towards quadrant 3 after 2050, combining worse economic performance and higher emissions. Limitations of the current modelling framework suggest that the analysis is likely to overstate the relative economic performance of the no-action scenarios (orange) and understate that of the very strong abatement scenarios (purple), because it does not fully account for all potentially significant climate impacts[1,2].

Making progress towards sustainable prosperity

In summary, we find that Australia could materially ease environmental pressures while enjoying strong economic growth. Many of the 20 scenarios we explored would represent substantial progress towards sustainable prosperity[46]. Australia could begin to repair past damage; restoring significant areas of native habitat and achieving negative emissions (net sequestration) of greenhouse gasses. But none of these scenarios would guarantee sustainability, or eliminate future threats to Australia's natural capital and the Earth's life-support systems[6,46]. Instead, each implies a different portfolio of risks and opportunities, which we have not fully modelled beyond 2050. For example, new native habitat established before 2050 could provide a permanent flow of biodiversity benefits and other ecosystem services, while the flow of carbon sequestration provided will peak and eventually decline to zero, drawing attention to challenges and opportunities beyond our modelling horizon, such as the possibility of using carbon plantations to generate negative emission bioenergy with carbon capture and storage[49].

Reducing environmental pressures will not require a shift in societal values, but neither will technology deliver it automatically. Collective choices and public policy settings have a crucial contribution, and well-designed markets can boost national income by exploiting new areas of comparative advantage in some circumstances. However, these scenarios may present new longer-term risks and opportunities, and the synergies and trade-offs involved will be influenced by global circumstances. We also find an important threshold effect: moderate global action to reduce greenhouse emissions may diminish Australia's traditional comparative advantage (particularly in fossil fuel-based sectors) without creating new areas of advantage; while stronger global action that places tangible value on emissions reductions could create new opportunities for creating value, providing win-win economic and environmental benefits relative to existing trends. While Australia could dramatically reduce environmental pressures across a wide range of global contexts, the economic costs of doing so will be smaller (and benefits larger) in global settings that support the stable functioning of key Earth systems, including through promoting clean energy. As these global circumstances emerge, Australia's opportunities will multiply.

Sustainable prosperity is possible, but not predestined. Australia is free to choose.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 22 May; accepted 9 October 2015.

1. Hatfield-Dodds, S. *et al. Australian National Outlook 2015: Living standards, resource use, environmental performance and economic activity, 1970–2050.* (CSIRO, Canberra, 2015).
2. Hatfield-Dodds, S. *et al. Australian National Outlook 2015 – Technical Report: Living standards, resource use, environmental performance and economic activity, 1970–2050.* (CSIRO, Canberra, 2015).

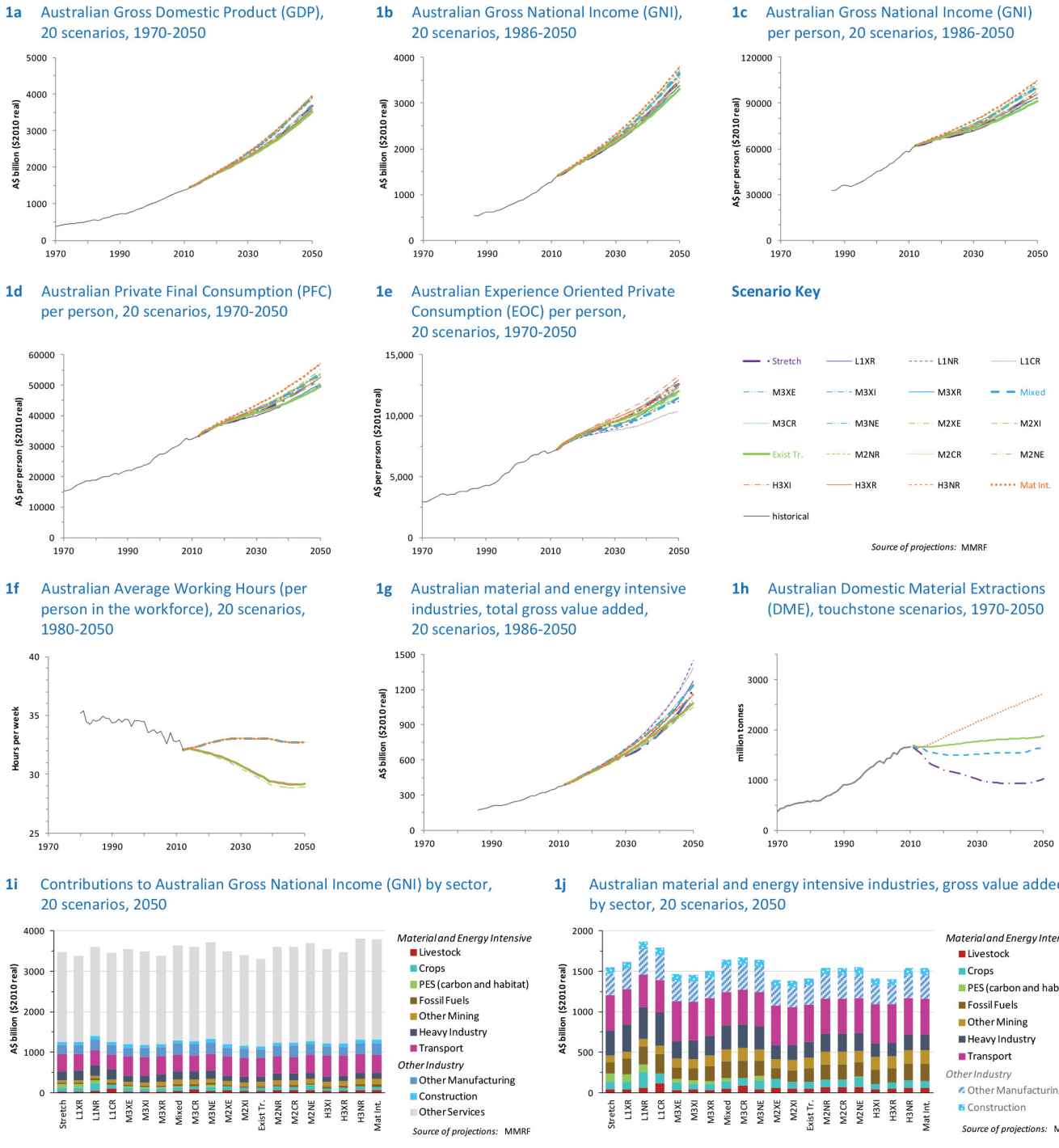
3. Lui, J. *et al.* Systems integration for global sustainability. *Science* **347,** 1–9 (2015).
4. Walker, B. *et al.* Looming global-scale failures and missing institutions. *Science* **325,** 1345–1346 (2009).
5. Stern, N. The structure of economic modelling of the potential impacts of climate change: grafting gross underestimation of risk onto already narrow science models. *J. Econ. Lit.* **51,** 838–859 (2013).
6. Steffen, W. *et al.* Planetary boundaries: guiding human development on a changing planet. *Science* **346,** 1–10 (2015).
7. World Bank. *Turn Down the Heat: Why a 4°C Warmer World Must be Avoided*. (World Bank, 2012).
8. Bryan, B. *et al.* Supply of carbon sequestration and biodiversity services from Australia's agricultural land under global change. *Glob. Environ. Change* **28,** 166–181 (2014).
9. UN, OECD, IMF, Eurostat World Bank (eds) *System of National Accounts 1993* (United Nations, Geneva, 1993).
10. Ferrier, S. *et al.* Mapping more of terrestrial biodiversity for global conservation assessment. *Bioscience* **54,** 1101–1109 (2004).
11. Ferrier, S., Manion, G., Elith, J. & Richardson, K. Using generalized dissimilarity modelling to analyse and predict patterns of beta diversity in regional biodiversity assessment. *Divers. Distrib.* **13,** 252–264 (2007).
12. Chiew, F. H. S. & Prosser, I. in *Water: Science and Solutions for Australia* (ed. Prosser, I.) 29–46 (CSIRO, Canberra, 2011).
13. Schandl, H. *et al.* Decoupling global environmental pressure and economic growth: scenarios for energy use, materials and carbon emissions. *J. Clean. Prod.* (2015).
14. Adams, P. D. & Parmenter, B. R. in *Handbook of Computable General Equilibrium Modelling* (eds Dixon, P. B. & Jorgenson, D. W.) Volume 1A. (Elsevier BV, 2013).
15. Graham, P. W. *et al. Modelling the Future Grid Forum Scenarios*. (CSIRO: Newcastle, 2013).
16. Jotzo, F. *et al.* in *Pathways to Deep Decarbonisation: 2014 Report*. (eds Guérin, E., Mas, C. & Waisman, H.) (Sustainable Development Solutions Network (SDSN) and Institute for Sustainable Development and International Relations (IDRRI), New York, 2014).
17. Graham, P., Brinsmead, T. & Hatfield-Dodds, S. Australian retail electricity prices: can we avoid repeating the rising trend of the past? *Energy Policy* **86,** 456–469 (2015).
18. UNEP. *Decoupling 2: Technologies, Opportunities and Policy Options*. (eds von Weizsäcker, E.U. *et al.*) 1–158 (United Nations Environment Program, Nariobi, 2014).
19. Wiedmann, T. O. *et al.* The material footprint of nations. reassessing resource productivity. *Proc. Natl Acad. Sci.* **112,** 6271–6276 (2015).
20. Fischer-Kowalski, M. *et al.* Methodology and indicators of economy wide material flow accounting: state of the art and reliability across sources. *J. Ind. Ecol.* **15,** 855–876 (2011).
21. Hejazi, M. I. *et al.* Integrated assessment of global water scarcity over the 21st century under multiple climate change mitigation policies. *Hydrol. Earth Syst. Sci.* **18,** 2859–2883 (2014).
22. DeFries, R. J. *et al.* Global nutrition: metrics for land-scarce agriculture. *Science* **349,** 238–240 (2015).
23. Hobday, A. J. & McDonald, J. Environmental issues in Australia. *Ann. Rev. Environ. Resour.* **39,** 1–28 (2014).
24. Decoupled ideals: 'Economodernist Manifesto' reframes sustainable development, but the goal remains the same. *Nature* **520,** 407–408 (2015).
25. Arrow, K. *et al.* Economic growth, carrying capacity, and the environment. *Science* **268,** 520–521 (1995).
26. Stern, D. I. The rise and fall of the environmental Kuznets curve. *World Dev.* **32,** 1419–1439 (2004).
27. Asafu-Adjaye, J. *et al. An Ecomodernist Manifesto*, 1–31 (2015). http://static1.squarespace.com/static/5515d9f9e4b04d5c3198b7bb/t/552d37bbe4b07a7dd69fcdbb/1429026747046/An+Ecomodernist+Manifesto.pdf.
28. Simon, J. L. *The Ultimate Resource*. (Princeton University Press, 1981).
29. Wildavsky, A. & Dake, K. Theories of risk perception: who fears what and why? *Daedalus* **19,** 41–60 (1990).
30. Kahan, D. Fixing the communications failure. *Nature* **463,** 296–297 (2010).
31. Ostrom, E. *Understanding Institutional Diversity*. (Princeton University Press, 2005).
32. Stern, N. The economics of climate change. *Am. Econ. Rev.* **98,** 1–37 (2008).
33. Lebel, L. *et al.* Governance and the capacity to manage resilience in regional social-ecological systems. *Ecol. Soc.* **11,** 19 (2006).
34. European Environment Agency. *Late lessons from early warnings: science, precaution, innovation*, EEA Report No 1/2013. (European Environmental Agency, Copenhagen, 2013).
35. Meadows, D. H., Meadows, D. L., Randers, J. & Behrens, W.W. *The limits to growth.* (Universe Books, New York, 1972).
36. Randers, J. *2052: A Global Forecast for the Next Forty Years.* (Chelsea Green Publishing, Vermont, 2012).
37. Rees, W. E. Achieving sustainability: reform or transformation? *J. Plann. Lit.* **9,** 343–361 (1995).
38. Daly, H. E. Economics in a full world. *Sci. Am.* **293,** 100–107 (2005).
39. Costanza, R. *et al. Building a Sustainable and Desirable Economy-in-Society-in-Nature*. (UN Division for Sustainable Development, New York, 2012).
40. Oliver, A. *The Lowy Institute Poll 2015*. (Lowy Institute for International Policy, Sydney, 2015).
41. Garnaut, R. *The Garnaut Climate Change Review: Final Report*. (Cambridge Uni. Press, Port Melbourne, 2008).
42. Garnaut, R. *The Garnaut Review 2011: Australia in the Global Response to Climate Change*. (Cambridge Uni. Press, Port Melbourne, 2011).
43. Climate Change Authority. *Reducing Australia's Greenhouse Gas Emissions – Targets and Progress Review: Final Report*. (Climate Change Authority, Melbourne, 2014).
44. Nordhaus, W. D. Economic aspects of global warming in a post-Copenhagen environment. *Proc. Natl Acad. Sci. USA* **107,** 11721–11726 (2010).
45. Dietz, T., Ostrom, E. & Stern, P. C. The struggle to govern the commons. *Science* **302,** 1907–1912 (2003).
46. Griggs, D. *et al.* Sustainable development goals for people and planet. *Nature* **495,** 305–307 (2013).
47. Haines, A. *et al.* Public health benefits of strategies to reduce greenhouse-gas emissions: overview and implications for policy makers. *Lancet* **374,** 2104–2114 (2009).
48. West, J. J. *et al.* Co-benefits of mitigating global greenhouse gas emissions for future air quality and human health. *Nature Clim. Change* **3,** 885–889 (2013).
49. Fuss, S. *et al.* Commentary: Betting on negative emissions. *Nature Clim. Change* **4,** 850–853 (2014).

Supplementary Information is available in the online version of the paper.

Acknowledgements The authors thank CSIRO Land and Water, CSIRO Energy, CSIRO Agriculture, and CSIRO Oceans and Atmosphere for funding and support, and J. Dowse of Clarity Thought Partners for assistance in preparing this paper and the National Outlook report.

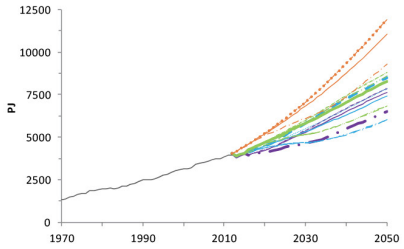
Author Contributions S.H-D. led the National Outlook project and oversaw all analysis, and led the drafting of this paper. All authors contributed to the analysis and interpretation, and commented on the draft paper, focusing as follows: S.H-D., study design, integration, and interpretation; H.S., material flows; P.D.A., CGE modelling; T.M.B., efficiency potential; T.S.B., transport; B.A.B. and M.N., land use; F.H.S.C. and I.P., water; P.W.G., stationary energy; M.G., agriculture; T.H., biodiversity; R.McCa., model linking, data integrity, analysis and charts; R.McCr., historical consumption trends; L.E.M., data integrity, analysis and charts, land and water analysis; D.N., global economics and climate; A.W., interpretation.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to S.H.-D. (Steve.Hatfield-Dodds@csiro.au).

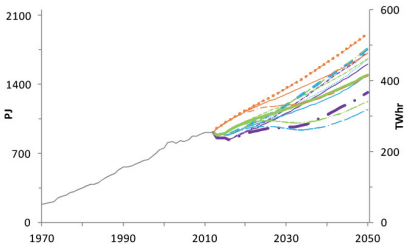


Extended Data Figure 1 | Australian economic activity, income and living standards, and material and energy intensive industries to 2050. Projections for 20 scenarios for nine indicators, and touchstone scenarios for one indicator. Income, consumption, and average working hours provide indicators of living standards. PES refers to payments for ecosystem services (carbon sequestration and habitat restoration). Definitions of scenarios and scenario assumptions, details of scenario sets, a full list of indicators, and references for historical data are provided in the Supplementary Information.

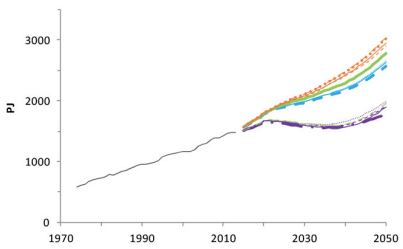
2a Australian Final Energy Demand, 20 scenarios, 1970-2050



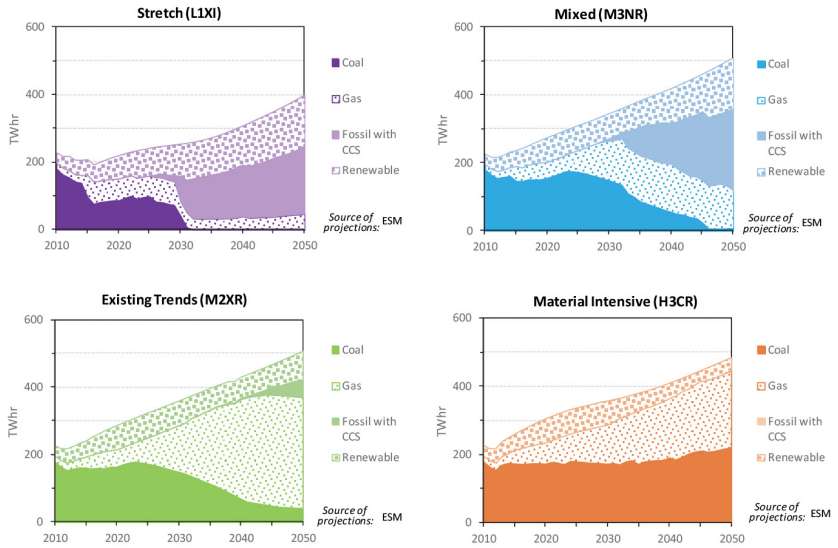
2b Australian Electricity Demand, 20 scenarios, 1970-2050



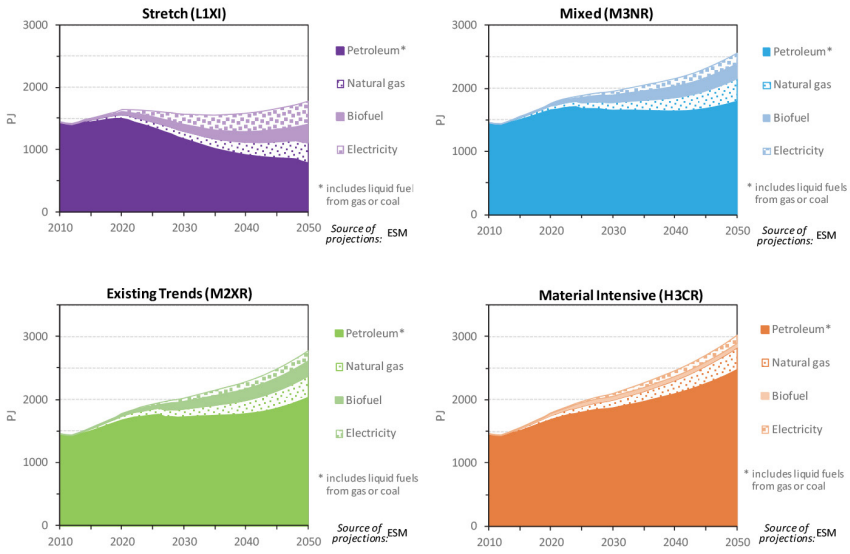
2c Australian Transport Energy Demand, 20 scenarios, 1970-2050



2d Australian Electricity Supply by Source, touchstone scenarios, 2010-2050

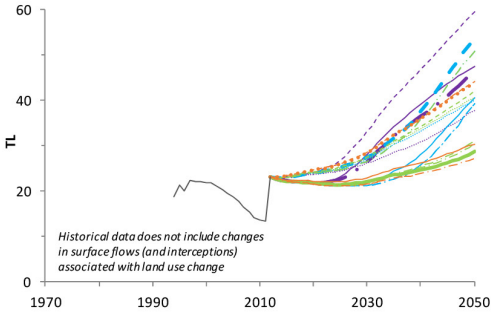


2e Australian Transport Energy by Fuel Type, touchstone scenarios, 2010-2050

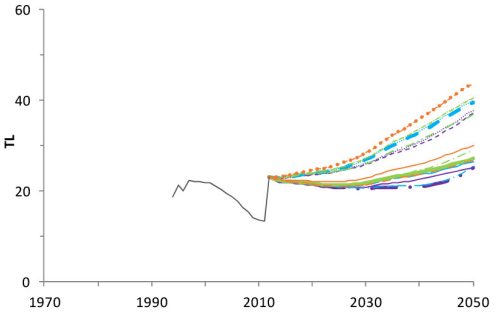


Extended Data Figure 2 | Australian energy use to 2050. Projections for 18 or 20 scenarios for three indicators, and touchstone scenarios for two indicators. Definitions of scenarios and scenario assumptions, details of scenario sets, a full list of indicators, and references for historical data are provided in the Supplementary Information. CCS, carbon capture and storage.

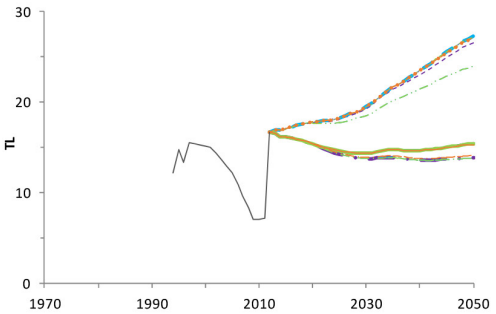
3a Water use, total (including interceptions), all catchments, 20 scenarios, 1994-2050



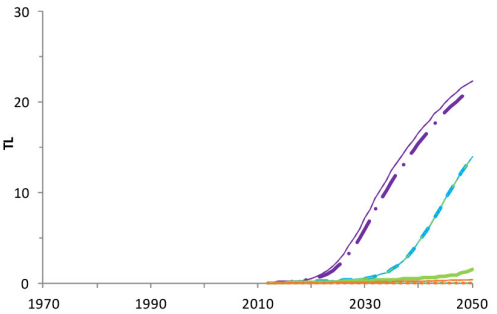
3b Extractive water use, all catchments, 18 scenarios, 1994-2050



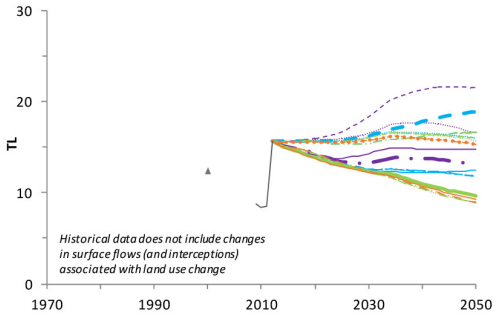
3c Agricultural extractive water use, all catchments, 18 scenarios, 1994-2050



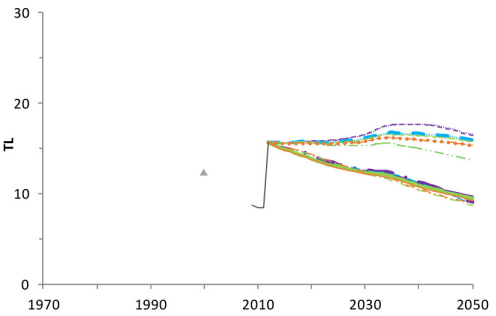
3d Water interceptions from land use change, all catchments, 18 scenarios, 2013-2050



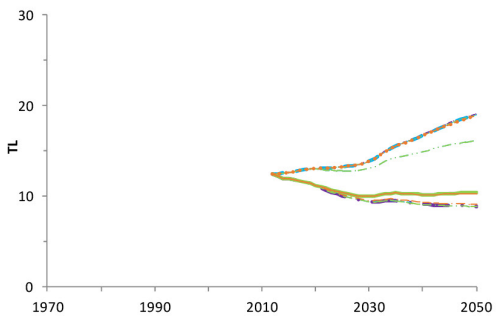
3e Water use, total (including interceptions), water limited catchments, 20 scenarios, 2000-2050



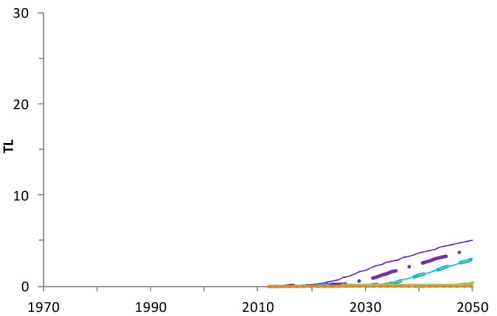
3f Extractive water use, water limited catchments, 18 scenarios, 2000-2050



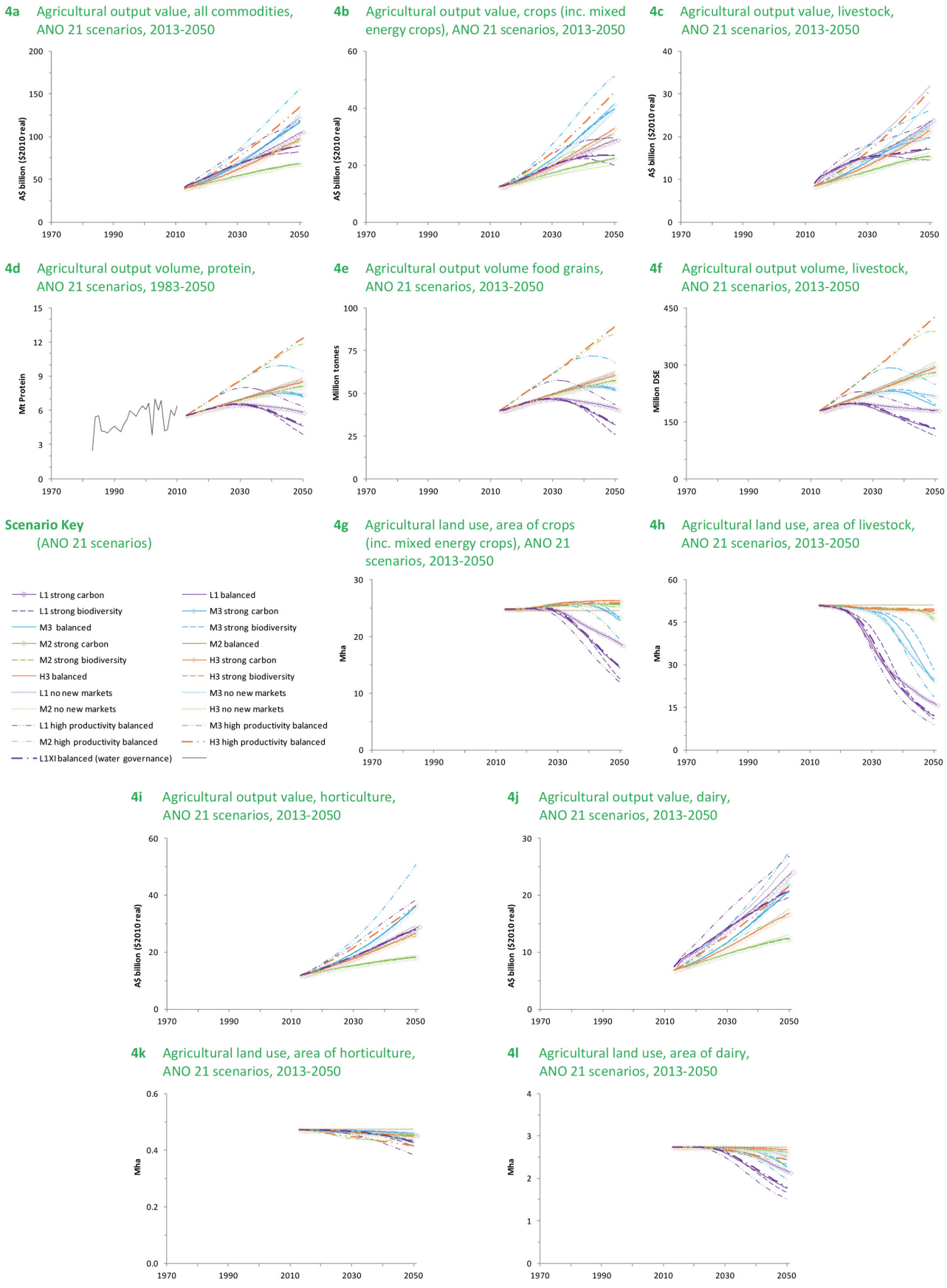
3g Agricultural extractive water use, water limited catchments, 18 scenarios, 2012-2050



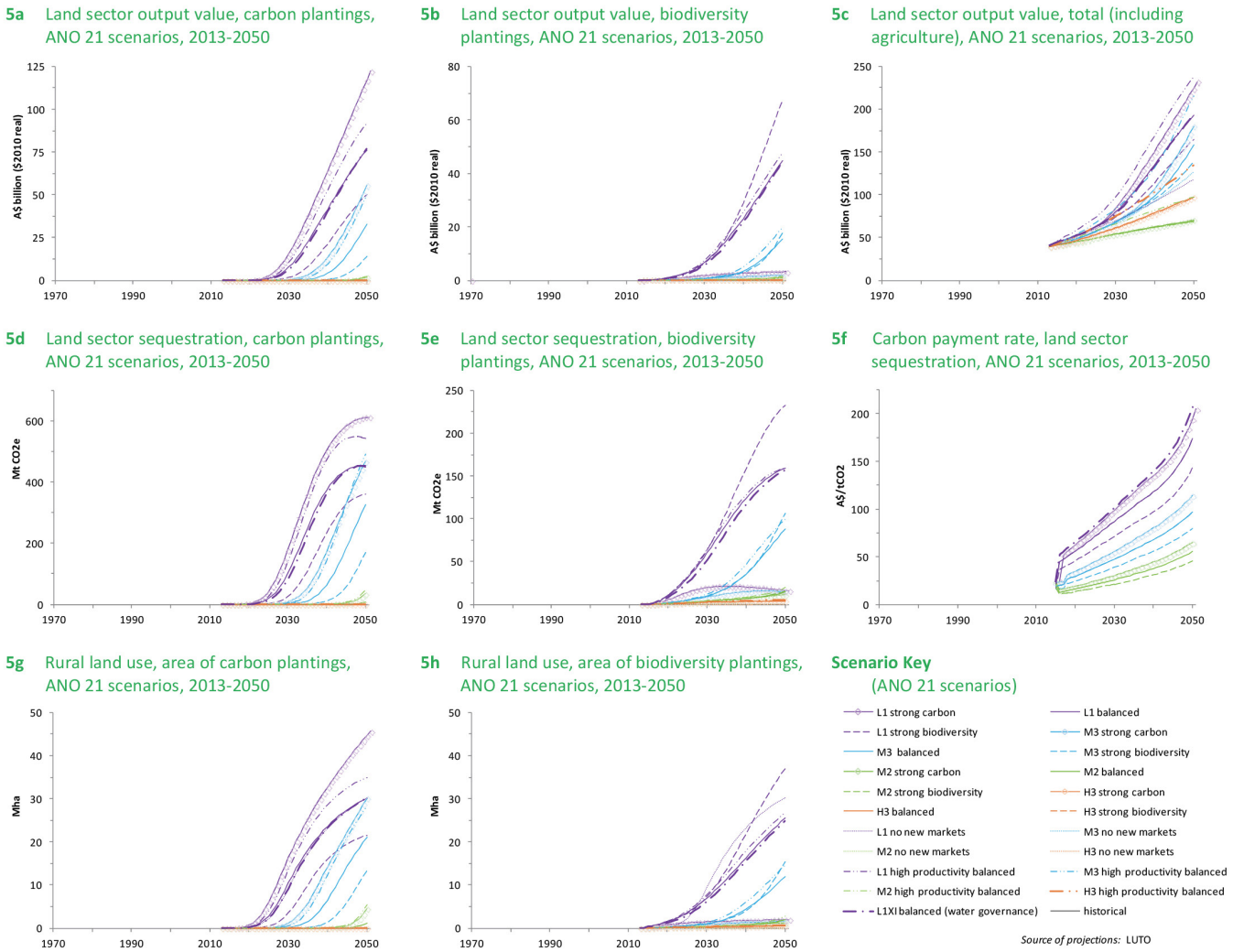
3h Water interceptions from land use change, water limited catchments, 18 scenarios, 2013-2050



Extended Data Figure 3 | Australia water use to 2050. Projections for 20 scenarios for two indicators and 18 scenarios for six indicators. Total water use is made up of extractive use plus interceptions of surface flows by new plantings that would otherwise contribute to streamflow. Water use in water-limited catchments provides an indication of water stress. Definitions of scenarios and scenario assumptions, details of scenario sets, a full list of indicators, and references for historical data are provided in the Supplementary Information.

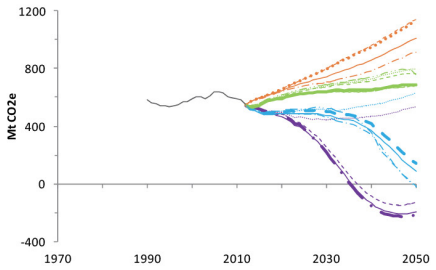


Extended Data Figure 4 | Australian agriculture output values, volumes and land use to 2050. Projections for 21 scenarios for 12 indicators. Food grains are a sub-set of crops. Protein calculation based on agricultural output volumes for all food commodities (including cereals, beef, sheep, legumes and dairy milk), weighted using USDA (2014). Definitions of scenarios and scenario assumptions, details of scenario sets, a full list of indicators, and references for historical data are provided in the Supplementary Information.

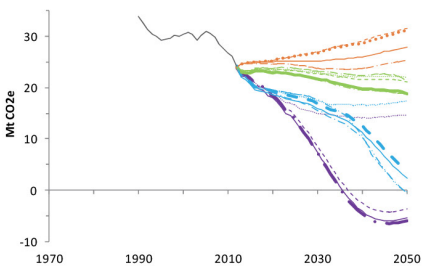


Extended Data Figure 5 | Australian land sector output values, volumes and land use to 2050. Projections for 21 scenarios for eight indicators. Total land sector activity is made up of agriculture (detailed in Extended Data Fig. 4) and payments for ecosystem services (carbon sequestration and habitat restoration) (see Extended Data Fig. 1i, j). Definitions of scenarios and scenario assumptions, details of scenario sets, a full list of indicators, and references for historical data are provided in the Supplementary Information.

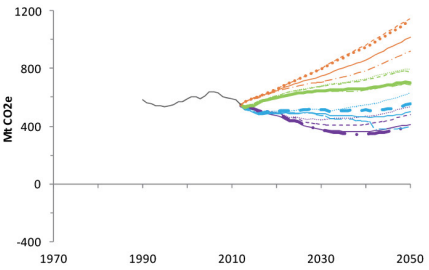
6a Domestic Net Greenhouse Gas Emissions, 18 scenarios, 1990-2050



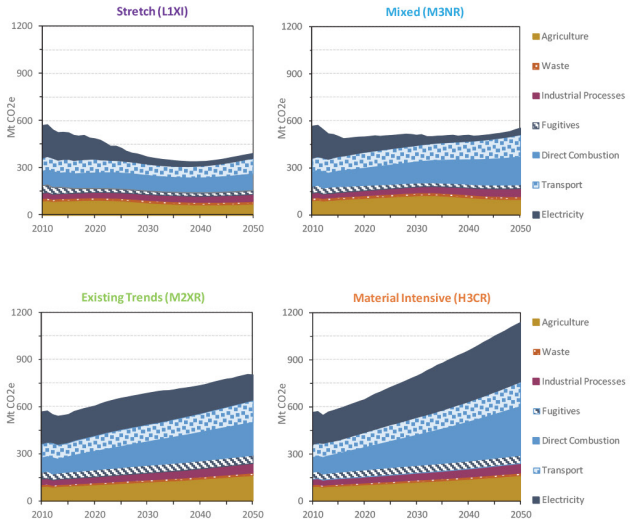
6b Domestic Net Greenhouse Gas Emissions per person, 18 scenarios, 1990-2050



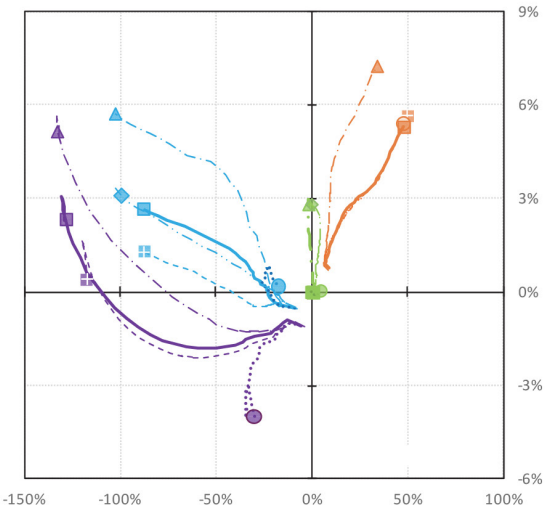
6c Domestic Greenhouse Gas Emissions (not including land sequestration), 18 scenarios, 1990-2050



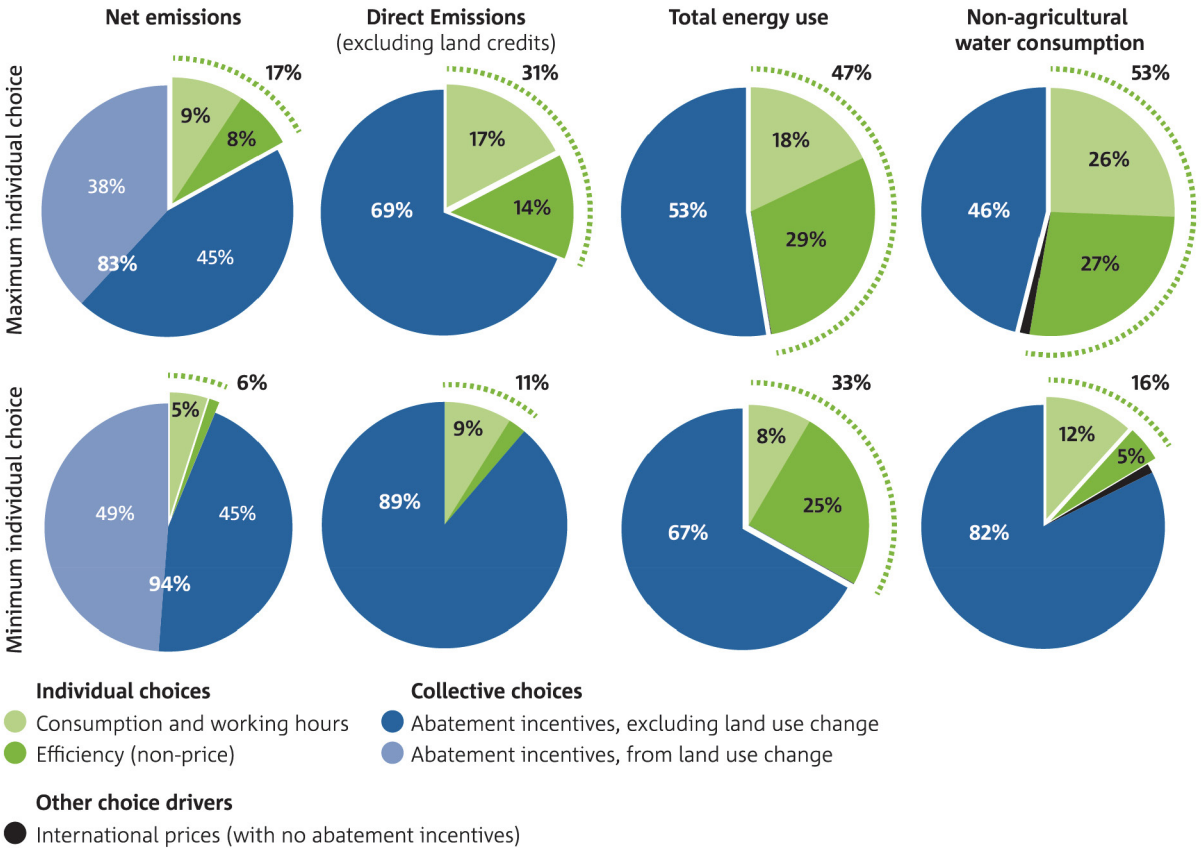
6d Australian Domestic Greenhouse Gas Emissions by source, not including land sector sequestration, touchstone scenarios, 2010-2050



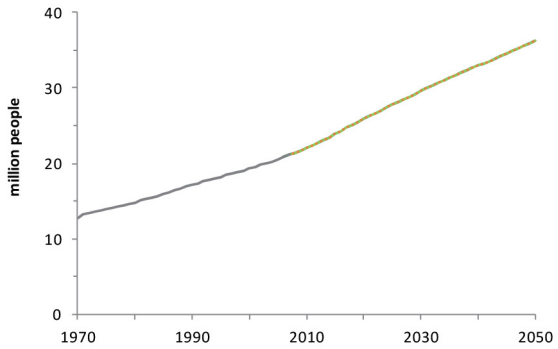
6e Deviation in Australian National Income (GNI) and Net Domestic Greenhouse Gas Emissions, 18 scenarios, 2015-2050



Extended Data Figure 6 | Australian greenhouse gas emissions and abatement to 2050. Projections for 18 scenarios for four indicators, and touchstone scenarios for one indicator. Domestic net emissions are defined as direct emissions less carbon sequestration (CCS and biosequestration) before trade in international emissions units. Calculations for Extended Data Fig. 6e are set out in Supplementary Methods, 'Calculations for Fig. 3 and assessment of potential economic performance with different levels of global and national action to reduce greenhouse emissions'. Definitions of scenarios and scenario assumptions, details of scenario sets, a full list of indicators, and references for historical data are provided in the Supplementary Information.

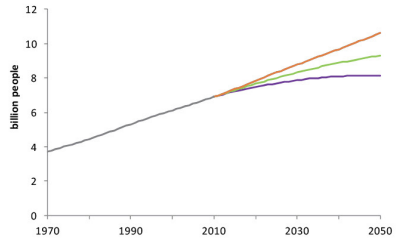


Extended Data Figure 7 | Maximum and minimum contributions of individual and collective choices to differences in projected greenhouse gas emissions, energy use, and non-agricultural water use in 2050. Calculations based on 20 scenarios, as described in Supplementary Methods, 'Assessing the contributions of individual and collective choices', drawing on data from Extended Data Figs 6a, b, 2a and 3b, c. Scenario assumptions and characteristics of the modelling framework prevent meaningful analysis of other indicators of environmental pressure for this purpose, such as total water use including agricultural extractions. Definitions of scenarios and scenario assumptions, details of scenario sets, a full list of indicators, and references for historical data are provided in the Supplementary Information.

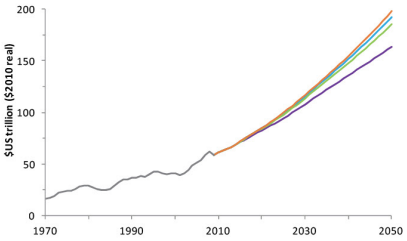


Extended Data Figure 8 | Australian population, 1970–2050. Population trajectory assumed in all domestic National Outlook scenarios. Information on age structure and dependency ratios is provided in ref 1. Definitions of scenarios and scenario assumptions, details of scenario sets, a full list of indicators, and references for historical data are provided in the Supplementary Information.

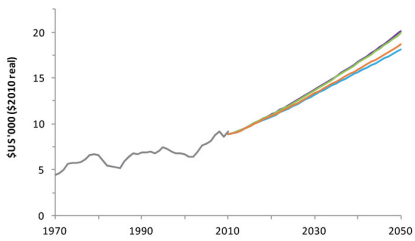
9a World population (scenario assumption), four global scenarios, 1970-2050



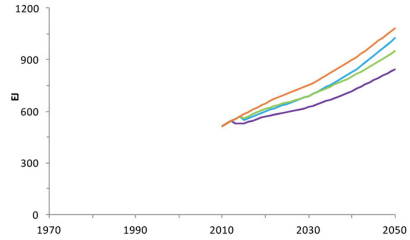
9b Gross World Product (GWP), four global scenarios, 1970-2050



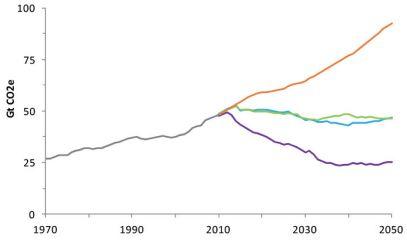
9c Gross World Product (GWP) per person, four global scenarios, 1970-2050



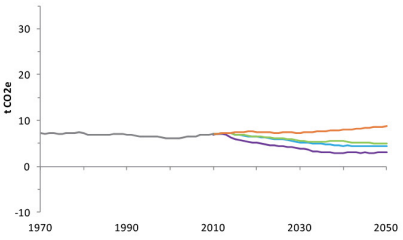
9d World Primary Energy Supply four global scenarios, 2010-2050



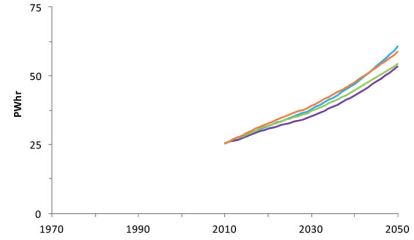
9e World Net Greenhouse Gas Emissions four global scenarios, 1970-2050



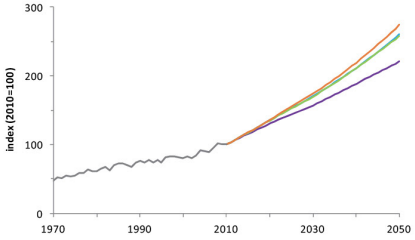
9f World Net Greenhouse Gas Emissions per person, four global scenarios, 1970-2050



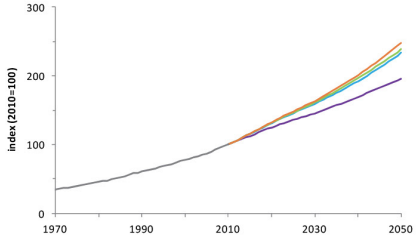
9g World Electricity Demand four global scenarios, 2010-2050



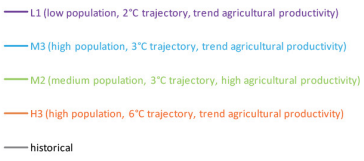
9h World Crops Output Volume, four global scenarios, 1970-2050



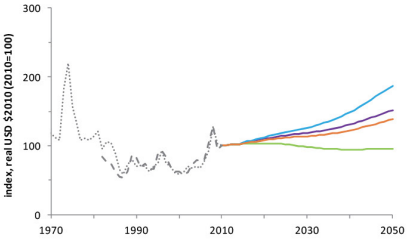
9i World Livestock Output Volume, four global scenarios, 1970-2050



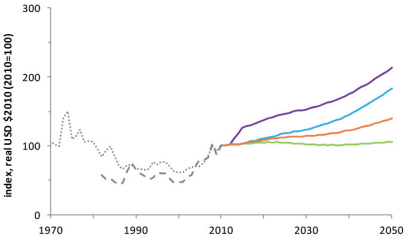
Scenario Key
(four global context scenarios)



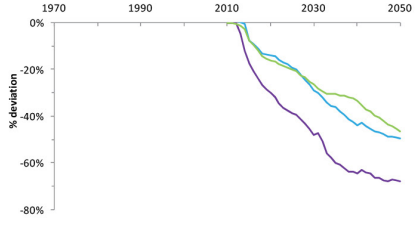
9j World Crops Price Index, four global scenarios, 1970-2050



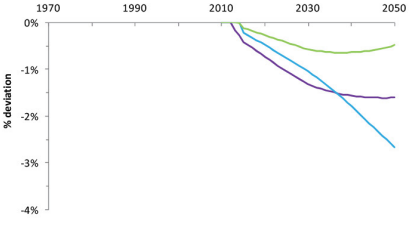
9k World Livestock Price Index, four global scenarios, 1970-2050



9l Deviation in World Net Greenhouse Gas Emissions, four global scenarios, 2010-2050



9m Deviation in Gross World Product (GWP), four global scenarios, 2010-2050



Note to ED-9m: Economic impact of abatement includes reduced agricultural production associated with reafforestation and avoided land clearing. This impact is proportionally larger per unit of abatement in scenarios with world higher population (blue) and lower for medium population (green) and low population (purple).

Extended Data Figure 9 | World population, economic activity, energy, emissions and agriculture to 2050. Projections for four global context scenarios for 11 indicators, and for three global context scenarios for two indicators. The global scenarios assume different combinations of population and cumulative greenhouse gas emissions, implying different levels of global abatement effort as well as different patterns of global demand and supply of energy and agricultural products. To give a wider range of contexts, the M2 (medium population, moderate abatement) global scenario also assumes higher global agricultural productivity, resulting in lower agricultural prices than would be projected otherwise. Definitions of scenarios and scenario assumptions, details of scenario sets, a full list of indicators, and references for historical data are provided in the Supplementary Information.

ARTICLE

doi:10.1038/nature15710

Combinatorial gene regulation by modulation of relative pulse timing

Yihan Lin[1,2], Chang Ho Sohn[3], Chiraj K. Dalal[1,2]†, Long Cai[3] & Michael B. Elowitz[1,2]

Studies of individual living cells have revealed that many transcription factors activate in dynamic, and often stochastic, pulses within the same cell. However, it has remained unclear whether cells might exploit the dynamic interaction of these pulses to control gene expression. Here, using quantitative single-cell time-lapse imaging of *Saccharomyces cerevisiae*, we show that the pulsatile transcription factors Msn2 and Mig1 combinatorially regulate their target genes through modulation of their relative pulse timing. The activator Msn2 and repressor Mig1 showed pulsed activation in either a temporally overlapping or non-overlapping manner during their transient response to different inputs, with only the non-overlapping dynamics efficiently activating target gene expression. Similarly, under constant environmental conditions, where Msn2 and Mig1 exhibit sporadic pulsing, glucose concentration modulated the temporal overlap between pulses of the two factors. Together, these results reveal a time-based mode of combinatorial gene regulation. Regulation through relative signal timing is common in engineering and neurobiology, and these results suggest that it could also function broadly within the signalling and regulatory systems of the cell.

In order to respond to environmental conditions, cells make extensive use of combinatorial gene regulation, in which two or more transcription factors co-regulate common target genes. Most analysis of combinatorial regulation presumes that the concentrations of transcription factors in the nucleus are regulated in a continuous (non-pulsatile) manner[1,2]. However, recent work has identified a large and growing list of transcription factors that activate in pulses[3–11]. In such systems, a single pulse begins when many molecules of a given transcription factor are activated simultaneously, and ends when they are deactivated. Such pulses can occur repetitively, even under constant conditions. Pulsatile regulation has been observed in bacteria[9,12,13], yeast[8,10,14–17], and mammalian stress response and signalling pathways[6,7,11,18–23]. In these systems, inputs typically modulate the pulse frequency, amplitude, and/or duration of individual transcription factors to regulate genes. However, despite analysis of many individual pulsatile transcription factors, the interactions between multiple pulsatile systems in the same cell have not yet been explored and analysed.

Saccharomyces cerevisiae provides an ideal model system to analyse such dynamic transcription factor interactions. It contains several well-characterized pulsatile systems that control core cellular functions. In particular, the general stress response transcription factor Msn2, and its paralogue, Msn4, activate hundreds of target genes in response to diverse stresses including ethanol, heat, oxidative stress, salt, and glucose starvation[24–30]. Similarly, the repressor Mig1, along with its paralogue, Mig2, control many target genes, especially those involved in metabolism, in response to changes in glucose concentration[31–33]. Together, Msn2 and Mig1 co-regulate over 300 target genes (according to Yeastract[34]). Both Msn2 and Mig1 are activated by dephosphorylation, which leads to nuclear localization[35–37]. Previous work has shown that Msn2 nuclear localization can occur in a pulsatile fashion in response to various inputs[8,10,14,17,38]. Mig1 is known to quickly localize to the nucleus in response to an increase in glucose levels[36], and can also exhibit pulsatile activation[38].

Two stages of dynamic pulsing

To analyse Msn2 and Mig1 dynamics in the same cell, we constructed strains expressing fusions of Msn2 and Mig1 proteins to the distinguishable fluorescent proteins[39] mKO2 and mCherry, respectively (Fig. 1a). To simplify the analysis, we knocked out their paralogues Msn4 and Mig2 (Methods). We attached single cells to the glass surface of a microfluidic channel, maintaining a constant flow of media, while acquiring time-lapse movies. By analysing individual cells in these movies, we could track the nuclear localization dynamics of both proteins over time (Methods).

We first analysed the effects of glucose reduction, which is known to induce changes in nuclear localization for both transcription factors[35,36]. In response to a sudden step from 0.2% to 0.1% glucose, both proteins exhibited pulses of nuclear localization, but did so with different timing (Fig. 1b). Msn2 localized to the nucleus immediately, while Mig1 exited the nucleus. Subsequently, in many cells (75%), Msn2 exited the nucleus followed by the re-entry of Mig1 (Fig. 1b; Supplementary Video 1). This transient response terminated within ~30 min (Fig. 1b, bottom). We describe events like this, in which Msn2 and Mig1 pulses are temporally separated, as non-overlapping (see Fig. 1b, top and Methods). After this event, Msn2 and Mig1 exhibited sporadic pulsing that was unsynchronized between cells (Supplementary Video 1). During this steady-state period, we observed both overlapping (that is, coincident) events, in which Msn2 and Mig1 pulses overlap, as well as non-overlapping events in which Msn2, but not Mig1 localized to the nucleus (Fig. 1b, top and Methods).

These data provoke two interrelated questions about whether and how relative pulse timing could function in combinatorial regulation (Fig. 1c): first, do inputs modulate the relative timing of transcription factor pulses, either during the transient response to a change in conditions, or during the subsequent period of repetitive pulsing? Second, if so, how does such pulse timing modulation affect downstream combinatorial gene regulation?

[1]Howard Hughes Medical Institute, California Institute of Technology, Pasadena, California 91125, USA. [2]Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, California 91125, USA. [3]Division of Chemistry and Chemical Engineering, California Institute of Technology, Pasadena, California 91125, USA. †Present address: Department of Microbiology and Immunology, UCSF, San Francisco, California 94143, USA.

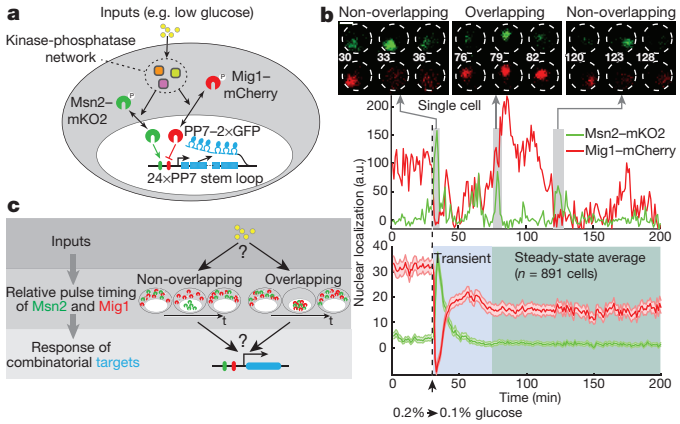


Figure 1 | Temporally structured pulsing of transcription factors Msn2 and Mig1 in response to glucose reduction. **a**, Inputs such as glucose regulate the phosphorylation and nuclear localization of Msn2 and Mig1, which co-regulate some common target genes. Three-colour strains allow simultaneous analysis of Msn2 and Mig1 nuclear localization dynamics and target gene expression. Yeast strains contained Msn2 (green) and Mig1 (red) fluorescent protein fusions, along with a target promoter with (shown) or without (not shown) binding sites for Msn2 and Mig1, driving expression of a transcript containing 24 stem-loops that are specifically bound by the PP7 RNA binding protein fused to 2 × GFP (blue circles). **b**, An example single-cell trace showing nuclear localization dynamics of Msn2 and Mig1. The cell exhibits an immediate temporally structured response to the step in glucose (arrowhead and dashed line), as well as sporadic pulsing throughout the movie. Filmstrips show examples of non-overlapping and overlapping events. White dashed circles indicate cell boundaries and numbers indicate time points. Scale bar is 2 μm. Lower plot shows average trace, revealing the synchronized transient non-overlapping response followed by a constant average response due to unsynchronized pulsing. Shading indicates 95% confidence intervals of the mean (Methods). **c**, These dynamics provoke the questions of how inputs modulate relative timing of Msn2 and Mig1 pulses, and how that timing affects gene regulation.

To address these questions, we constructed strains containing synthetic target promoters incorporating binding sites for either or both transcription factors (Fig. 1a). These promoters drove expression of a transcriptional reporter consisting of 24 binding sites for a separately expressed PP7 RNA binding protein fused to green fluorescent protein (GFP)[40] (Fig. 1a). These strains enabled us to simultaneously follow localization dynamics of Msn2 and Mig1 and downstream target expression in the same cell.

Relative pulse timing in the transient response

We first analysed transient responses to changes in various input conditions (that is, different Msn2 stressors) other than the known common input glucose (Fig. 2a). Addition of 100 mM NaCl produced transient non-overlapping pulses of Msn2 and Mig1 in single cells and in population averages (Fig. 2b, Extended Data Fig. 1a–c, Supplementary Video 2) that were similar to those observed in the transient response to glucose reduction (Fig. 1b). Addition of 2.5% ethanol also activated both transcription factors. But in contrast to NaCl, it did so with overlapping, rather than non-overlapping, pulses (Fig. 2c, Extended Data Fig. 1d–f, Supplementary Video 3). The difference in relative timing between NaCl and ethanol was also apparent in cross-correlation analysis (Extended Data Fig. 1g). Together, these results indicate that distinct inputs can generate opposite relative timing in the transient responses of Msn2 and Mig1.

We hypothesized that control of temporal overlap could provide a mechanism for combinatorial gene regulation. Non-overlapping pulse dynamics, in which the activator Msn2 is active, but the repressor Mig1 is not, could activate combinatorial target genes more efficiently than overlapping pulses, in which the two proteins are simultaneously bound to the same target promoter. Indeed, while both NaCl and ethanol led to activation of an Msn2-specific target

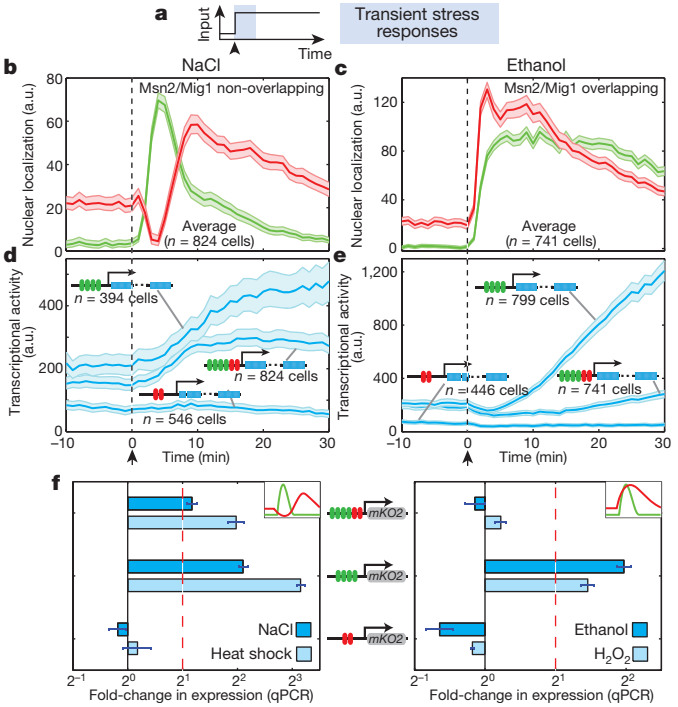


Figure 2 | Different inputs produce distinct transient gene expression responses by modulating relative pulse timing. **a**, Transient nuclear localization and gene expression responses were simultaneously monitored in individual cells. **b**, **c**, Addition of NaCl (100 mM) or ethanol (2.5%) induced non-overlapping and overlapping responses, respectively. Green and red traces show mean Msn2 and Mig1 nuclear localization, respectively. **d**, **e**, Averaged single-cell transcriptional activity traces show that NaCl activated both combinatorial and Msn2-specific targets, while ethanol activated only the Msn2-specific target. Shading in **b**–**e** indicates 95% confidence interval of the mean. **f**, qPCR data are consistent with single-cell data (**b**, **c**), and extend these responses to heat shock and H_2O_2 stresses (Extended Data Fig. 1h, i; see Methods). Error bars indicate s.e.m. calculated from 3–8 biological replicates.

promoter, only the non-overlapping dynamics of NaCl efficiently induced target expression (Fig. 2d, e, Extended Data Fig. 1a–f). Moreover, we observed similar timing-mediated regulation with other stresses. Heat shock and oxidative stress (from H_2O_2) induced non-overlapping and overlapping dynamics, respectively (Extended Data Fig. 1h, i). As with the other stresses, both non-overlapping and overlapping dynamics activated an Msn2-specific target promoter, but only non-overlapping dynamics efficiently activated the combinatorial target promoter (Fig. 2f). As expected, the dependence of expression from the synthetic combinatorial target promoter on relative timing required both Msn2 and Mig1 (Extended Data Fig. 2a). In addition, these effects were not specific to the synthetic target promoter, as expression of *GSY1* (ref. 41), an endogenous target of Msn2 and Mig1, exhibited similar dependence on relative timing in response to stresses, as shown by both single-cell analysis and quantitative PCR data (Extended Data Fig. 2b–e). In fact, further genome-wide analysis revealed 30 additional endogenous targets that exhibited a similar pattern of gene regulation during transient responses to NaCl and ethanol (Methods, Extended Data Fig. 2f–k, and Supplementary Discussion), suggesting that relative timing-dependent regulation applies to multiple endogenous target genes, as well as to the synthetic promoter. Together these data indicate that, during transient stress responses, cells regulate gene expression by modulating the relative pulse timing between Msn2 and Mig1.

Regulation by relative pulse timing at steady-state

We next asked whether relative pulse timing could also function in constant environmental conditions where both transcription factors

pulse sporadically and repetitively. Because such pulsing is not synchronized among cells, it could only be analysed with single-cell movie data. We observed both overlapping and non-overlapping pulse events under constant conditions (Fig. 1b, Fig. 3a, Extended Data Fig. 3a, b, and Supplementary Videos 4, 5). To better understand the effects of each type of event on gene expression, we adapted the technique of pulse-triggered averaging from neurobiology (usually called spike-triggered averaging)[42] (Extended Data Fig. 3c). We identified Msn2 pulses, and sorted them into two groups depending on whether or not a Mig1 pulse overlapped temporally with the Msn2 pulse (Fig. 3a, Methods). We then averaged the Msn2 and Mig1 dynamics over a time window around the Msn2 pulse peaks, for both overlapping and non-overlapping events. By construction, the resulting pulse-triggered averages showed opposite overall dynamic relationships between the two proteins (Fig. 3b, c).

Pulse-triggered averaging enabled us to analyse the dependence of target gene expression on Msn2 pulsing and, more specifically, on its temporal relationship with Mig1, averaged over variability in both pulsing behaviour and downstream transcriptional responses (see Supplementary Discussion about the multiple layers of variability in this system). Both overlapping and non-overlapping pulses led to subsequent increase in the mean expression of the pure Msn2 synthetic target promoter (Extended Data Fig. 4a–c). However, only the non-overlapping events showed activation of the synthetic combinatorial Msn2-Mig1 promoter or the natural combinatorial target gene, *GSY1* (Fig. 3d, e). Moreover, deletions of the zinc-finger DNA binding domains of either Msn2 or Mig1 eliminated the relative timing-dependence of *GSY1* expression, indicating that DNA-binding of both proteins is necessary for relative timing-dependent regulation (Extended Data Fig. 4d). Together, these results show that relative timing between Msn2 and Mig1 pulses regulates gene expression under steady-state conditions.

Thus far, we have simplified the analysis of relative pulse timing by classifying events as either overlapping or non-overlapping. However, cross-correlation analysis revealed more complexity in the dynamics. For example, we observed a peak at a positive time lag of ~2–4 min, corresponding to sequential activation of Msn2 followed by Mig1 (Extended Data Fig. 4f–i, also evident in Fig. 3c, f; see Supplementary Discussion). More generally, the data showed a continuous distribution of time intervals between a given Msn2 pulse and its previous, or subsequent, Mig1 pulse. To better understand how these dynamics affect target gene expression, we analysed the dependence of mean expression level on the continuous time interval between Msn2 and Mig1 pulses (Extended Data Fig. 5a, b). Mean gene expression is minimal when Msn2 and Mig1 pulse simultaneously, but Mig1 pulses occurring within ~ 4–5 min before or after Msn2 pulses also suppress mean expression. These results are consistent with a model in which Mig1 pulses can both terminate continuing expression from preceding Msn2 pulses, and also establish promoter states with reduced tendency to activate in response to Msn2, possibly due to residual binding of Mig1 itself or to Mig1-induced effects on promoter states. As expected, these extended timing effects required both Msn2 and Mig1 binding sites on the target promoter, as well as DNA-binding activities of both proteins (Extended Data Fig. 5d, e). These characteristic timescales for Msn2-Mig1 pulse interactions establish the degree of simultaneity necessary for pulses to function as overlapping events.

Modulation of relative pulse timing

Having established the effect of relative pulse timing on gene expression at steady-state, we next asked whether and how inputs affect relative timing. We acquired time-lapse movies of Msn2 and Mig1 nuclear localization across a range of constant glucose concentrations (from 0.4% to 0.0125%), where both Msn2 and Mig1 exhibited sporadic nuclear localization pulses (Extended Data Fig. 6, 7b–e). The frequencies of pulses for both proteins, and the mean duration of Mig1 pulses, all varied systematically with glucose concentration (Extended Data Fig. 7a), while mean

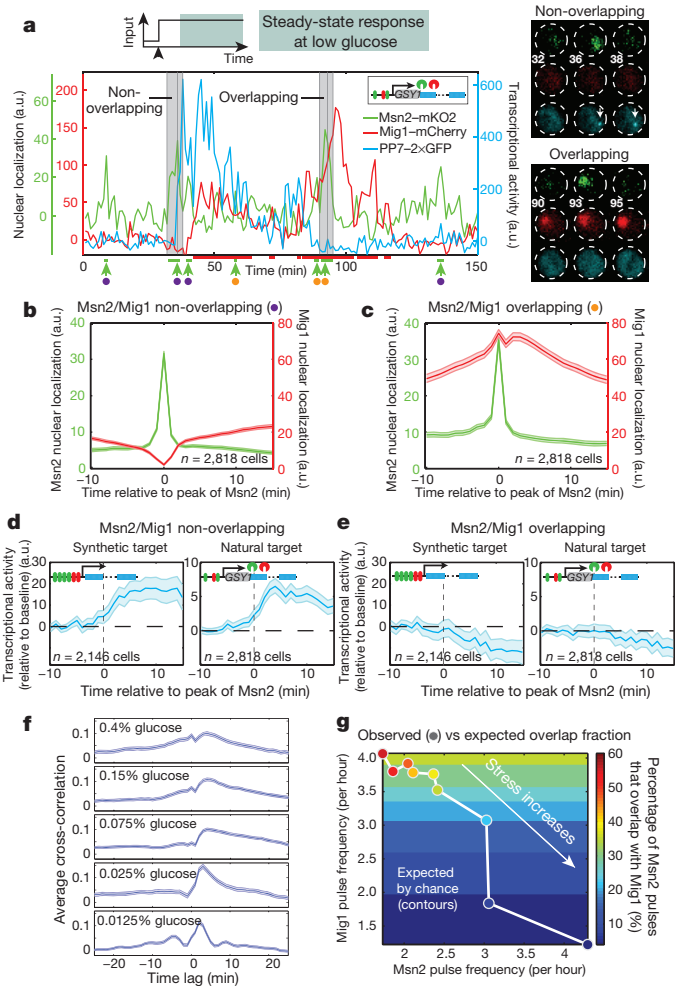


Figure 3 | Pulse-triggered averaging reveals relative pulse timing-dependent gene expression under constant conditions, and modulation of relative timing by glucose concentration. **a**, Localization and target transcription dynamics in a single cell under constant (0.05%) glucose. Msn2 and Mig1 localization are shown in green and red, respectively, while transcriptional activity of their co-regulated target, *GSY1* (*GSY1-24xPP7SL*) is shown in blue. Filmstrips show examples of non-overlapping and overlapping events (indicated by grey shading). White arrows on the upper filmstrip indicate active transcriptional sites for the target gene. Green and red horizontal lines below plot indicate identified Msn2 and Mig1 pulses. Green arrows indicate peaks of the Msn2 pulses used for pulse-triggered averaging (Methods). **b**, **c**, Pulse-triggered averages of Msn2 and Mig1 localization events sorted into non-overlapping (**b**, purple; $n = 14{,}384$ events) and overlapping (**c**, orange; $n = 7{,}829$ events) groups. **d**, **e**, Pulse-triggered average transcriptional activity traces for non-overlapping (**d**) and overlapping (**e**) events. Baseline activity (horizontal dashed line) was subtracted from each trace. Traces are aligned to the peak Msn2 pulse at $t = 0$ (vertical dashed line). **f**, Cross-correlation between Msn2 and Mig1 dynamics at different glucose levels (see also Extended Data Fig. 7g). **g**, Glucose levels modulate the percentage of Msn2 pulses that overlap with Mig1. Circles indicate measurements of pulse frequency (location of circle) and the percentage of Msn2 pulses that overlap with Mig1 (overlap fraction, colour of circle) for nine glucose levels (from 0.4% to 0.0125% as in Extended Data Fig. 8a). Horizontal contours indicate the overlap fraction expected at each glucose level assuming independent Msn2 and Mig1 dynamics (Methods). See also controls in Extended Data Fig. 8f, g. Shading indicates 95% confidence intervals of the mean.

pulse amplitudes remained approximately constant (Extended Data Fig. 7a). Interestingly, however, averaged cross-correlations between Msn2 and Mig1 nuclear localization traces showed features (for example, the peak at time lag zero) that depended on glucose concentration (Fig. 3f). Furthermore, the percentage of Msn2 pulses that overlap with Mig1,

which we define as the overlap fraction, changed systematically with glucose concentration (Fig. 3g and Extended Data Fig. 8a). Together, these results indicate that glucose concentration modulates the relative pulse timing between Msn2 and Mig1 at steady-state conditions.

To better understand the effect of glucose concentration on relative pulse timing, it is helpful to distinguish between passive and active types of modulation. Passive modulation arises from changes in the frequency and/or duration of Mig1 pulses, and occurs even if Msn2 and Mig1 dynamics are independent. By contrast, active modulation would require mechanisms that specifically enhance or reduce the fraction of overlapping events.

Passive modulation seems to dominate at lower glucose concentration, but both passive and active modulation occur at higher glucose concentrations. At very low glucose levels (<0.05%), the observed overlap fraction agreed with expectations based on passive modulation only (Methods, lower right of Fig. 3g and Extended Data Fig. 8a). However, at higher glucose levels (≥0.05%), where pulse frequencies became less glucose-dependent (Extended Data Fig. 7a), the observed overlap fraction exceeded the value expected from passive modulation, and increased systematically with glucose concentration (upper left corner of Fig. 3g and Extended Data Fig. 8a), indicating a substantial role for active modulation. Moreover, including the active component of modulation improved the ability of a simple model to explain the dependence of target gene expression on glucose (Extended Data Fig. 8b–d and Supplementary Discussion). We also found that relative pulse timing could be further modulated by other inputs such as NaCl and ethanol (Extended Data Fig. 9 and Supplementary Discussion). These results show that, under steady-state conditions, input identity (type of stress) and level (for example, glucose concentration) together modulate relative pulse timing, through both passive and active mechanisms, to control target gene expression.

Mechanism for relative pulse timing modulation

Relative pulse timing modulation represents a distinct mode of gene regulation that operates in both steady-state and transient conditions (Fig. 4a, see also Supplementary Discussion). What mechanisms could enable cells to actively control relative pulse timing? One possibility involves regulatory components that specifically generate overlapping pulses of Msn2 and Mig1. Previous work has shown that Glc7, the catalytic component of PP1 phosphatase, can indirectly regulate both Msn2 and Mig1 nuclear localization[43], making it a candidate for an active regulator of overlapping pulses (Extended Data Fig. 10a). We constructed a strain in which the wild-type *GLC7* promoter was replaced with a Cu^{2+}-inducible promoter in the native locus. In this strain, reducing expression of *GLC7* below wild-type levels abolished active modulation, making the measured overlap fraction equal to that expected by chance (overlap of red solid and dashed lines in the left panel of Fig. 4b). This effect can also be seen in the Msn2-Mig1 cross-correlation at time lag zero, which is reduced at higher glucose concentrations (compare red and black lines in Fig. 4b, right). Restoring *GLC7* expression close to wild-type levels restored active modulation (blue lines, Fig. 4b). Together, these data (Fig. 4b and Extended Data Fig. 10) support a role for Glc7 in active modulation by glucose (Supplementary Discussion). Other phosphoregulatory components may also contribute to active modulation in these and other conditions.

Discussion

What functions could relative pulse timing modulation provide for the cell? One of the most fundamental concepts in combinatorial regulation is that cooperative interactions between transcription factors can increase their probability of simultaneous binding to a promoter, to implement *cis*-regulatory logic[44]. By controlling the fraction of time that two transcription factors are simultaneously active, relative pulse timing modulation could provide similar effects in *trans*

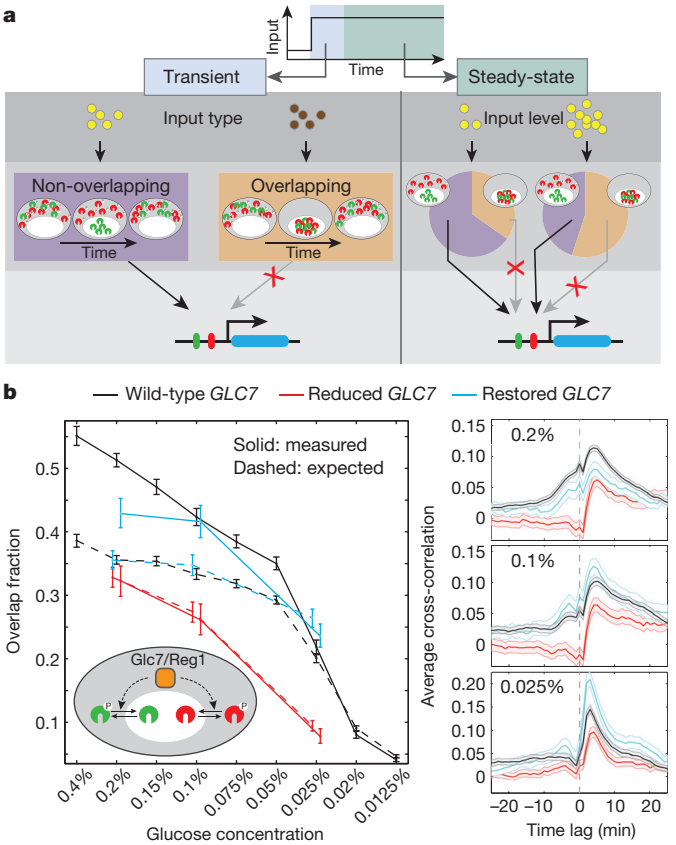


Figure 4 | Mechanistic aspect of relative pulse timing modulation. **a**, In gene regulation by relative pulse timing modulation (schematic), the identity and level of inputs (yellow and brown circles) regulate target gene expression through changes in the relative timing of Msn2 and Mig1 pulses (see Supplementary Discussion). Overlapping events (orange) only activate Msn2-specific targets while non-overlapping events (purple) activate both Msn2-specific and Msn2-Mig1 combinatorial targets. In steady-state (right), inputs modulate the fraction of Msn2 pulses that overlap with Mig1 (pie charts). **b**, *GLC7* mediates active modulation of relative pulse timing, possibly by activating both Msn2 and Mig1 (schematic inset, left). Left, measured (solid line) and expected (dashed line) overlap fractions were plotted for three conditions: wild-type (black), reduced *GLC7* expression (red), and the same strain with *GLC7* expression restored to approximately wild-type levels (blue). Right, average cross-correlation between Msn2 and Mig1 dynamics for three glucose levels (percentages). Shading and error bars indicate 95% confidence intervals of the mean.

(Supplementary Note and Extended Data Fig. 10f–h). In addition to its functionality, a number of basic issues about timing-dependent regulation remain to be understood. For example, what accounts for variability among cells in their transcription factor dynamics and the apparently stochastic response of target promoters to those dynamics? What features of target promoters, such as the kinetic parameters that govern their activation, determine whether and how they respond to timing-based regulation?

Relative timing between signals plays many important roles throughout science and engineering. In neuroscience, the relative timing of action potentials at pre- and post-synaptic neurons controls the strength of synaptic connectivity through spike-timing-dependent plasticity[45]. In communications, modulating the phase of a periodic signal relative to a reference signal is widely used to encode information[46]. Cells seem to have evolved a related strategy by encoding aspects of the extracellular environment in the relative timing with which different transcription factors pulse. The unsynchronized nature of these pulses (at steady-state) has made relative pulse timing modulation rather difficult to detect and characterize previously. However, pulsatile dynamics (both periodic and aperiodic) are now being

discovered in a growing list of central signalling and regulatory pathways[3–5], which are known to interact, or crosstalk, with one another. It will therefore be critical to more systematically map the temporal organization of cellular pathways, and determine principles that can explain both the mechanisms and functions of relative pulse timing modulation in living cells.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 2 January; accepted 4 September 2015.
Published online 14 October 2015.

1. Alon, U. *An Introduction to Systems Biology: Design Principles of Biological Circuits.* (Taylor & Francis, 2006).
2. Gertz, J., Siggia, E. D. & Cohen, B. A. Analysis of combinatorial *cis*-regulation in synthetic and genomic promoters. *Nature* **457,** 215–218 (2009).
3. Yosef, N. & Regev, A. Impulse control: temporal dynamics in gene transcription. *Cell* **144,** 886–896 (2011).
4. Purvis, J. E. & Lahav, G. Encoding and decoding cellular information through signaling dynamics. *Cell* **152,** 945–956 (2013).
5. Levine, J. H., Lin, Y. & Elowitz, M. B. Functional roles of pulsing in genetic circuits. *Science* **342,** 1193–1200 (2013).
6. Lahav, G. *et al.* Dynamics of the p53-Mdm2 feedback loop in individual cells. *Nature Genet.* **36,** 147–150 (2004).
7. Nelson, D. E. *et al.* Oscillations in NF-κB signaling control the dynamics of gene expression. *Science* **306,** 704–708 (2004).
8. Cai, L., Dalal, C. K. & Elowitz, M. B. Frequency-modulated nuclear localization bursts coordinate gene regulation. *Nature* **455,** 485–490 (2008).
9. Locke, J. C., Young, J. W., Fontes, M., Hernández Jiménez, M. J. & Elowitz, M. B. Stochastic pulse regulation in bacterial stress response. *Science* **334,** 366–369 (2011).
10. Hao, N. & O'Shea, E. K. Signal-dependent dynamics of transcription factor translocation controls gene expression. *Nature Struct. Mol. Biol.* **19,** 31–39 (2012).
11. Cohen-Saidon, C., Cohen, A. A., Sigal, A., Liron, Y. & Alon, U. Dynamics and variability of ERK2 response to EGF in individual living cells. *Mol. Cell* **36,** 885–893 (2009).
12. Young, J. W., Locke, J. C. & Elowitz, M. B. Rate of environmental change determines stress response specificity. *Proc. Natl Acad. Sci. USA* **110,** 4140–4145 (2013).
13. Levine, J. H., Fontes, M. E., Dworkin, J. & Elowitz, M. B. Pulsed feedback defers cellular differentiation. *PLoS Biol.* **10,** e1001252 (2012).
14. Garmendia-Torres, C., Goldbeter, A. & Jacquet, M. Nucleocytoplasmic oscillations of the yeast transcription factor Msn2: evidence for periodic PKA activation. *Curr. Biol.* **17,** 1044–1049 (2007).
15. Hao, N., Budnik, B. A., Gunawardena, J. & O'Shea, E. K. Tunable signal processing through modular control of transcription factor translocation. *Science* **339,** 460–464 (2013).
16. Hansen, A. S. & O'Shea, E. K. Promoter decoding of transcription factor dynamics involves a trade-off between noise and control of gene expression. *Mol. Syst. Biol.* **9,** 704 (2013).
17. Petrenko, N., Chereji, R. V., McClean, M. N., Morozov, A. V. & Broach, J. R. Noise and interlocking signaling pathways promote distinct transcription factor dynamics in response to different stresses. *Mol. Biol. Cell* **24,** 2045–2057 (2013).
18. Kholodenko, B. N., Hancock, J. F. & Kolch, W. Signalling ballet in space and time. *Nature Rev. Mol. Cell Biol.* **11,** 414–426 (2010).
19. Tay, S. *et al.* Single-cell NF-κB dynamics reveal digital activation and analogue information processing. *Nature* **466,** 267–271 (2010).
20. Batchelor, E., Loewer, A., Mock, C. & Lahav, G. Stimulus-dependent dynamics of p53 in single cells. *Mol. Syst. Biol.* **7,** 488 (2011).
21. Albeck, J. G., Mills, G. B. & Brugge, J. S. Frequency-modulated pulses of ERK activity transmit quantitative proliferation signals. *Mol. Cell* **49,** 249–261 (2013).
22. Yissachar, N. *et al.* Dynamic response diversity of NFAT isoforms in individual living cells. *Mol. Cell* **49,** 322–330 (2013).
23. Kageyama, R., Ohtsuka, T., Shimojo, H. & Imayoshi, I. Dynamic Notch signaling in neural progenitor cells and a revised view of lateral inhibition. *Nature Neurosci.* **11,** 1247–1251 (2008).
24. Martínez-Pastor, M. T. *et al.* The *Saccharomyces cerevisiae* zinc finger proteins Msn2p and Msn4p are required for transcriptional induction through the stress response element (STRE). *EMBO J.* **15,** 2227–2235 (1996).
25. Schmitt, A. P. & McEntee, K. Msn2p, a zinc finger DNA-binding protein, is the transcriptional activator of the multistress response in *Saccharomyces cerevisiae*. *Proc. Natl Acad. Sci. USA* **93,** 5777–5782 (1996).
26. Boy-Marcotte, E., Perrot, M., Bussereau, F., Boucherie, H. & Jacquet, M. Msn2p and Msn4p control a large number of genes induced at the diauxic transition which are repressed by cyclic AMP in *Saccharomyces cerevisiae*. *J. Bacteriol.* **180,** 1044–1052 (1998).
27. Estruch, F. Stress-controlled transcription factors, stress-induced genes and stress tolerance in budding yeast. *FEMS Microbiol. Rev.* **24,** 469–486 (2000).
28. Gasch, A. P. *et al.* Genomic expression programs in the response of yeast cells to environmental changes. *Mol. Biol. Cell* **11,** 4241–4257 (2000).
29. Hasan, R. *et al.* The control of the yeast H_2O_2 response by the Msn2/4 transcription factors. *Mol. Microbiol.* **45,** 233–241 (2002).
30. Morano, K. A., Grant, C. M. & Moye-Rowley, W. S. The response to heat shock and oxidative stress in *Saccharomyces cerevisiae*. *Genetics* **190,** 1157–1195 (2012).
31. Nehlin, J. O., Carlberg, M. & Ronne, H. Control of yeast GAL genes by MIG1 repressor: a transcriptional cascade in the glucose response. *EMBO J.* **10,** 3373–3377 (1991).
32. Lutfiyya, L. L. & Johnston, M. Two zinc-finger-containing repressors are responsible for glucose repression of SUC2 expression. *Mol. Cell. Biol.* **16,** 4790–4797 (1996).
33. Carlson, M. Glucose repression in yeast. *Curr. Opin. Microbiol.* **2,** 202–207 (1999).
34. Teixeira, M. C. *et al.* The YEASTRACT database: an upgraded information system for the analysis of gene and genomic transcription regulation in *Saccharomyces cerevisiae*. *Nucleic Acids Res.* **42,** D161–D166 (2014).
35. Görner, W. *et al.* Nuclear localization of the C2H2 zinc finger protein Msn2p is regulated by stress and protein kinase A activity. *Genes Dev.* **12,** 586–597 (1998).
36. De Vit, M. J., Waddle, J. A. & Johnston, M. Regulated nuclear translocation of the Mig1 glucose repressor. *Mol. Biol. Cell* **8,** 1603–1618 (1997).
37. Treitel, M. A., Kuchin, S. & Carlson, M. Snf1 protein kinase regulates phosphorylation of the Mig1 repressor in *Saccharomyces cerevisiae*. *Mol. Cell. Biol.* **18,** 6273–6280 (1998).
38. Dalal, C. K., Cai, L., Lin, Y., Rahbar, K. & Elowitz, M. B. Pulsatile dynamics in the yeast proteome. *Curr. Biol.* **24,** 2189–2194 (2014).
39. Shaner, N. C., Steinbach, P. A. & Tsien, R. Y. A guide to choosing fluorescent proteins. *Nature Methods* **2,** 905–909 (2005).
40. Larson, D. R., Zenklusen, D., Wu, B., Chao, J. A. & Singer, R. H. Real-time observation of transcription initiation and elongation on an endogenous yeast gene. *Science* **332,** 475–478 (2011).
41. Unnikrishnan, I., Miller, S., Meinke, M. & LaPorte, D. C. Multiple positive and negative elements involved in the regulation of expression of *GSY1* in *Saccharomyces cerevisiae*. *J. Biol. Chem.* **278,** 26450–26457 (2003).
42. Meister, M., Pine, J. & Baylor, D. A. Multi-neuronal signals from the retina: acquisition and analysis. *J. Neurosci. Methods* **51,** 95–106 (1994).
43. De Wever, V., Reiter, W., Ballarini, A., Ammerer, G. & Brocard, C. A dual role for PP1 in shaping the Msn2-dependent transcriptional response to glucose starvation. *EMBO J.* **24,** 4115–4123 (2005).
44. Ptashne, M. *A Genetic Switch: Phage Lambda Revisited* (Cold Spring Harbor Laboratory Press, 2004).
45. Bi, G. Q. & Poo, M. M. Synaptic modifications in cultured hippocampal neurons: dependence on spike timing, synaptic strength, and postsynaptic cell type. *J. Neurosci.* **18,** 10464–10472 (1998).
46. Anderson, J. B., Aulin, T. & Sundberg, C.-E. *Digital Phase Modulation* (Springer, 1986).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank U. Alon, R. Corral, R. Deshaies, A. Eldar, J. Garcia-Ojalvo, R. Kishony, A. Moses, G. Seelig, P. Swain, and members of the Elowitz laboratory for comments and feedback on the manuscript. We also thank the core sequencing facility at Caltech for help on RNA-Seq. This work was supported by the NIH (R01 GM079771B, R01 GM086793A), the NSF (Award no. 1547056), DARPA (HR0011-05-1-0057), and by the Gordon and Betty Moore Foundation through Grant GBMF2809 to the Caltech Programmable Molecular Technology Initiative. L.C. acknowledges the Ellison foundation for support.

Author Contributions Y.L. and M.B.E. designed experiments. Y.L. performed experiments and analysed data with input from all authors. C.K.D. and L.C. initially observed the correlation between Msn2 and Mig1 dynamics and C.H.S. conducted preliminary analysis of target gene expression. M.B.E. supervised research. Y.L. and M.B.E. wrote the manuscript with input from all authors.

Author Information RNA-Seq data have been deposited at Gene Expression Omnibus (GEO) under the accession code GSE71712. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to M.B.E. (melowitz@caltech.edu).

METHODS

Strain construction. Standard protocols were used for molecular cloning. Plasmids were replicated in either TOP10 or DH5α *Escherichia coli*. Except where indicated, all yeast strains used in this study were constructed based on BY4741 (*MATa his3Δ0 leu2Δ0 met15Δ0 ura3Δ0*), where *msn4*, *mig2*, *nrg1*, *nrg2* were further deleted (seamless deletion) or compromised (with auxotrophic or drug markers) to avoid complications resulting from these proteins binding to Msn2 or Mig1 binding sites. All yeast transformations were performed with standard lithium-acetate protocol[47] or with Frozen-EZ Yeast Transformation II Kit (Zymo Research). Resulting constructs were confirmed with PCR and/or sequencing. Details of strain genotypes are listed in Supplementary Table 1.

For endogenous gene fusion, *MSN2-mKO2::LEU2* and *MIG1-mCherry::spHIS5* were constructed by the fusion PCR approach where a PCR product comprised of 300–500 bp of 3′ end of target of interest, *mKO2* or *mCherry* gene, *LEU2* or sp*HIS5* cassette, and another 300–500 bp of the target downstream. More specifically, *mCherry*::sp*HIS5* was directly PCR amplified from pKT355 plasmid, *mKO2* gene was obtained from Amalgaam Co., Ltd, and *LEU2* was amplified from pRS315 plasmid. Fused PCR products were directly transformed. For RNA binding protein fusion PP7–2 × GFP, pDZ276 plasmid (a gift from R. Singer) was directly used for transformation into yeast.

Synthetic promoters driving either *24xPP7SL* binding cassette (for single-cell 3-colour movies) or *mKO2* (for qPCR measurements) are composed of the following elements: *ADH1* terminator–UAS–basal *HIS3* promoter (−101 to −1 of *HIS3* gene)–*24xPP7SL* cassette with *ADH1* terminator or *mKO2–KANMAX* or *NATMX* resistance cassette. *ADH1* terminator and *KANMX* cassette were obtained from pKT vectors[48]. *NATMX* was obtained from pAG25 plasmid[49]. Basal *HIS3* promoter was amplified from yeast genome. The *24xPP7SL* cassette was obtained from Addgene (plasmid 31864). mKO2 was used for qPCR analysis because it is exogenous to yeast genome. Three different UAS cassettes contained one or both of the following elements: 4 copies of Msn2 binding motif (GATCTACAGCCCCTGGAAAAT, adopted from *HSP12* promoter[26]) and/or 2 copies of Mig1 binding motif (AATAAAAATGCGGGGAA, adopted from *SUC2* promoter[50]). These UAS cassettes were used to generate Msn2-specific, Mig1-specific, and Msn2-Mig1 combinatorial promoters. The entire constructs were flanked with sequences for integration into *TRP1* locus of BY4741 and were assembled into a pKT based vector. The plasmids were digested with AfeI to release the entire cassette for integration into respective yeast strains. *GSY1-24xPP7SL* (for 3-colour movies) was generated by integration of *24xPP7SL::KANMX* cassette directly downstream of the endogenous *GSY1* gene.

Zinc finger deletion mutants of Msn2 and Mig1 proteins were constructed by direct transformation of PCR fragments containing desired mutations. Specifically, a fused PCR product containing *MIG1(Δamino acid36-91)-mCherry::spHIS5* was used to generate Mig1–mCherry with its DNA binding domain deleted. Similarly, a fused PCR product containing *MSN2(Δaa642-704)-mKO2::LEU2* was used for Msn2 zinc finger mutation. Deletion of Mig1 zinc finger appeared to impact its regulation of nuclear localization as the mutated Mig1–mCherry became much more nuclear localized. This effect, however, does not affect our conclusions.

Copper-inducible *GLC7* strain was constructed by transforming a fusion PCR product of URA3-TEF terminator-*CUP1* promoter flanked with sequences for integration to replace the endogenous *GLC7* promoter. Transformants were selected on plates containing 100 μM $CuSO_4$.

Media and growth conditions. We adopted a minimal media formula with low auto-fluorescence for both culturing yeast cells and for microscopy[8]. Stock solutions for minerals (1,000×), vitamins (1,000×), as well as salts (50×) were made separately. Final working media was made by mixing these three components together with amino acid drop-out mix (from Clontech) and Milli-Q water. Media was adjusted to desired glucose concentration with a glucose stock (40%, w/v).

For overnight liquid culture, single colonies of yeast were picked from agar plates made with minimal media and dispensed into 2–3 ml of minimal media (2% glucose, −Ura or −His −Leu −Ura) in 14 ml round-bottom polypropylene tubes (BD catalogue no. 352059). Cells were grown in a 30 °C shaking incubator. The media and overnight culture procedures were the same for both single-cell microscopy and qPCR experiments. For microscopy, media was supplemented with 2 mM sodium ascorbate (Sigma catalogue no. A7631) and 200 μM trolox (Sigma catalogue no. 238812) (except for media with H_2O_2) to help reduce fluorescent protein photobleaching and photo toxicity to cells.

Time-lapse microscopy. All time-lapse experiments were performed on an Olympus IX81 microscope with 60× objective and hardware autofocus (ZDC2). Fluorescence was excited by a LED light source (Lumencor SOLA Light Engine) and collected onto a scientific CMOS camera (Andor Neo sCMOS) with a 2-by-2 bin setting. For mKO2 and mCherry, single *z*-plane images were acquired. For GFP, a 5-slice *z*-stack was acquired (0.8 μm separation). The excitation and emission filters for mKO2, mCherry and GFP are: Ex 534/20 and Em 572/28, Ex 580/20 and Em 630/60, and Ex 472/30 and Em 535/50, respectively. The frame rate is 1 frame min^{-1}. Time-lapse movie automation was performed with Micro-Manager[51]. The entire microscope room was maintained at ~ 26 °C with two heater fans and a temperature controller (Omega Engineering catalogue no. FCH-FGC20012R and catalogue no. CSC32J).

Movies were acquired for single cells cultured in a dual-inlet microfluidic channel (~500 μm wide), which enables media switching. The microfluidic device was fabricated with polydimethylsiloxane (PDMS) with a Sylgard 184 silicone elastomer kit (DOW Corning) and bonded with 24 mm × 50 mm glass coverslip (Gold Seal No. 1.5) after air-plasma cleaning (Harrick Plasma PDC-32G). The channels were cleaned by brief incubation with 2 M NaOH, followed by washes with 100% ethanol and water. A 15 mg ml^{-1} concanavalin A (Sigma catalogue no. C7275) solution was incubated in the channels for about 10 min to coat the surface for adhering single yeast cells. Channels were washed with media before cell loading. Overnight yeast cultures were diluted back to $OD_{600\ nm} = 0.1$ with 2 ml of fresh media (0.2% glucose, −Ura) and were allowed to grow for another ~3 h. Cells were briefly concentrated by centrifuge and loaded into the channel. Cells were incubated in the channels for 5 min. The device was then loaded onto a sample stage on the microscope. Two inlets of a channel were connected with tubing (Weico Wire&Cable catalogue no. TT-30) to two different media solutions in 10 ml syringes (BD catalogue no. 309604) containing different glucose or stimulant concentrations. These syringes were driven with separate syringe pumps (Harvard Apparatus Pump 11 elite) which were controlled by Micro-Manager. Outlet of the channel was connected to a waste container. Media flow rate was maintained at 5 μl min^{-1} throughout the movie except for during media change (at 50 μl min^{-1} for 2 min).

The starting glucose concentration and the time before media switching differed in different experiments. For the transient glucose shift experiment (Fig. 1b), cells were in the channel with flowing 0.2% glucose media for more than 2 h before switching to 0.1% glucose (acquisition of fluorescent images started 30 min before switching). For experiments in Fig. 2, cells were in the channel with flowing 0.05% glucose media for more than 2 h before switching to 0.05% glucose plus specified stressor. For steady-state experiments in Figs 3, 4, cells were in the channel with flowing 0.2% glucose media for at least 10 min before switching to 0.05% or other designated glucose levels (from 0.4% to 0.0125%). Acquisition of fluorescent images started 110 min after switching media conditions (that is, at steady-state). For cooper inducible *GLC7* experiments (Fig. 4b), cells were cultured with minimal media without the addition of cooper until they were switched to a media containing 10 μM $CuSO_4$ for 110 min before the acquisition of fluorescent images.

Image analysis for extracting single-cell traces. Single-cell traces were extracted from fluorescence images based on cell tracking performed on bright-field images. All analyses were implemented with custom Matlab code (with some modules obtained online as cited below). More specifically, a slightly defocused bright-field image was taken at each frame for segmentation and tracking purposes. Segmentation was performed by circular Hough transformation (CircularHough_Grd function from Mathworks File Exchange). Segmented cell masks were first aligned across the entire movie frames to roughly correct for *x*-*y* stage drift (in order to enhance tracking accuracy). The masks were then fed into a tracking algorithm (u-track[52]) to obtain final cell tracks. Tracks were examined manually, and those with errors were discarded and removed from further analysis. These filtered single-cell tracks were used to extract fluorescence traces.

For analysis of, *z*-stack GFP images (for real-time transcription) we used a maximal intensity *z*-projection. Fluorescent images were then background subtracted (using background images acquired with media only) and corrected for field flatness caused by uneven illumination (using an image taken with fluorescein). Nuclear localization was calculated by the difference between the mean intensity of the top five pixels and the median intensity of all pixels. Single-cell nuclear localization traces for mKO2 and mCherry were then obtained with tracks obtained above. Real-time transcriptional activity (that is, PP7–2 × GFP signal) was calculated as the intensity of the brightest pixel in the cell minus its local background. For the time when transcription is active, the brightest pixel coincides well with the transcription hotspot. Nuclear localization and transcriptional activity measurements were validated by manual examination of the extracted traces side-by-side with fluorescence images.

Single-cell trace analysis and pulse-triggered averaging analysis. Single-cell traces were first baseline-subtracted and nuclear localization pulses were identified. These pulses were then characterized and used for pulse-triggered averaging analysis. More specifically, calculation of the baselines for mKO2 and mCherry traces was based on a measure for the degree of nuclear localization. In this method, pairwise spatial distance summed over the top 10 brightest pixels in individual cells was used to determine when the fluorescence signal is nuclear

localized or not at a given frame. Nuclear localization scores from frames with the summed distance above a predefined threshold were used to estimate the baseline by a polynomial fit. For cases in which the baseline varied too much along a trace, baseline was estimated by fitting only the nuclear localization values that were below an empirically defined threshold. Baseline for GFP signal was estimated by polynomial fitting the GFP signals that were below an empirically defined threshold. Baseline subtraction procedures were validated by manual examination.

Nuclear localization pulses were identified in both Msn2 and Mig1 traces. Pulse identification was based on iPeak (from Mathworks File Exchange). Shoulder peaks were filtered out and combined with neighbouring peaks (with higher amplitude). The remaining peaks were filtered based on an amplitude threshold (at least 20% above the baseline values) as well as the summed pairwise distance (below a predefined threshold). Width of the pulses was measured for left and right portions of the pulses separately (first fitted with spline and then measured at half of the pulse amplitude or the amplitude threshold, whichever is smaller). For pulse-triggered averaging analysis, a 21 min window around the peak of each Msn2 pulse (that is, 10 min on each side) was used for classifying the relative timing. This time window was chosen based on the frequency of Msn2 pulses. Within this window, all Mig1 pulses were identified. If the peak of a triggered Msn2 pulse fell into the span (defined by pulse width) of a Mig1 pulse, it was classified as an overlapping event. Otherwise, it was classified as a non-overlapping event. A more detailed classification based on the distance between the peak of Msn2 pulse and the edge (defined by pulse width) of the Mig1 pulse (if multiple Mig1 pulses occur within the window, the one with maximum pulse amplitude was chosen for this classification) can also be done as shown in Extended Data Fig. 5. Overlapping and non-overlapping events were averaged separately. Note that a larger time window (that is, a 26 min window with 10 min on the left of the peak and 15 min on the right) was chosen for averaging in Fig. 3 and Extended Data Figs 4, 5 in order to capture and measure the prolonged transcriptional responses in the GFP dynamics.

Cross-correlation analysis. Several figures include cross-correlation analysis (Figs 3f, 4b and Extended Data Figs 1g, 7g and 9d). In these cases, we first compute the cross-correlation function for each cell in a given data set, and then average the resulting functions. Individual cross-correlations were based on mean-subtracted signals and normalized, computed using the following expression:

$$C_{xy}(\tau)=\frac{\langle(x(t)-\langle x\rangle)\cdot(y(t+\tau)-\langle y\rangle)\rangle}{\sqrt{\langle(x(t)-\langle x\rangle)^2\rangle\langle(y(t)-\langle y\rangle)^2\rangle}}$$

Here, angled brackets denote means, and $C_{xy}(\tau)$ is the cross-correlation of $x(t)$ and $y(t)$ at time lag τ.

Quantitative PCR analysis. We used qPCR analysis to validate the single-cell transcriptional response, with similar culture procedures for both microscopy and RNA analysis. In this protocol, cells were exposed to defined stimulants for 10 min and RNA was extracted for two-step RT–qPCR (reverse transcription followed by qPCR). Note that the concentrations of salt, ethanol, and H_2O_2 were doubled when compared to the microfluidic single-cell assay (that is, 200 mM versus 100 mM NaCl, 5% ethanol versus 2.5% ethanol, 0.5 mM versus 0.25 mM H_2O_2). Overnight cultures were diluted to $OD_{600\,nm}=0.075$ with 20 ml of 0.2% glucose (−Ura) in 250 ml flask and allowed to grow until the $OD_{600\,nm}$ reached above 0.2 (about 3–4 h). For transient stress experiments, cultures were then diluted back to $OD_{600\,nm}=0.2$ with 20 ml of 0.05% glucose in 250 ml flask and allowed to grow for another 2 h. Cultures were split into 14 ml polypropylene tubes (4 ml each). Stresses were applied by mixing concentrated stock solutions (such as 5 M NaCl, 100% ethanol, 0.83 M H_2O_2) with the culture or by moving the culture tubes to a 37 °C shaking incubator (for heat shock). After precisely 10 min of stress application, each culture was mixed with 6 ml pre-chilled methanol (with dry ice/ethanol bath) in a 50 ml tube to rapidly fix the cells. For steady-state experiments, cells were diluted to $OD_{600\,nm}=0.1$ with 4 ml fresh media of designated condition (different glucose concentration with or without additional $CuSO_4$) in 50 ml falcon tube. Cultures were allowed to grow for 2 h and cells were mixed quickly with cold methanol as above. After >1 h in cold methanol, cells were collected by centrifuging at 4 °C and washed with ice cold water. Prior to performing standard RNA extraction protocols (with on-column DNase digestion) with RNeasy mini kits (Qiagen), cells were enzymatically treated with 100 µl of 2 U μl^{-1} lyticase solution (Sigma catalogue no. L2524) for 10 min at 30 °C. The extracted RNA absorbance spectrum was analysed with NanoDrop and 1 µg RNA was used for a standard 20 µl iScript (Bio-Rad) reverse transcription reaction. The resulting cDNA was diluted 4× with water before proceeding to qPCR reaction. A typical 10 µl qPCR reaction was assembled with 5 µl iQ SYBR Green Supermix (Bio-Rad), 2 µl primers (1.5 µM each), 2 µl of cDNA, and 1 µl of water. Reactions were performed on a CFX96 Real-Time machine (Bio-Rad). Each reaction had ≥2 technical replicates. Three reference genes were included (*ACT1, UBC6, TFC1*) for each sample. The latter two were based on recommendations by Teste *et al.*[53]. The mean Cq values of these reference genes were used for the calculation of ΔΔCq (or fold-change as $2^{-\Delta\Delta Cq}$) for each gene between sample and control. Calculations of ΔΔCq were done by CFX Manager Software (Bio-Rad) and final processing was performed by Matlab (Mathworks). Error bars were calculated by taking the standard errors of ≥3 biological replicates. Primers were designed according to manufacture instructions for iQ SYBR Green and were blasted against the yeast transcriptome (Primer-Blast[54]) to avoid nonspecific priming. The following primer sequences were used:

ACT1_F: ACATCGTTATGTCCGGTGGT; ACT1_R: CATGGAAGATGGAGCCAAAG; UBC6_F: AGGACCTGCGGATACTCCTT; UBC6_R: TCTGATAGCCGGTGGTTTGT; TFC1_F: AGCGCTGGCACTCATATCTT; TFC1_R: TTGGGCGTATTCCACTGAAC; mKO2_F: GTGATCAAGCCCGAGATGAA; mKO2_R: CATCTCCTGATGTCCCTCGT; GSY1_F: ACTGGTTGATTGAGGGAGCA; GSY1_R: GACCATAGGTCAGCCTTCCA; EMI2_F: AATGGTGACGGAACCTTTGA; EMI2_R: GCGACCCAGGTAGCTAAACA; GLC3_F: CCGCTCCATAGGTGGTACTG; GLC3_R: ACTTCCCATCTCCCATTCATC; GPH1_F: TCTGGCCACCCATGAATTAG; GPH1_R: GCAACGCTCAGGACACTCTT; IGD1_F: AGCAATGGTAACAGCGCAAG; IGD1_R: CTCCAAACATGTGAAGCTGGT.

RNA-Seq library construction and data analysis. For data shown in Extended Data Fig. 2, RNA-Seq was performed with libraries prepared from the RNA samples collected from cells of three different strains (no deletion strain and deletions of either *msn2* or *mig1*) subjected to no treatment (control), 200 mM NaCl, or 2.5% ethanol. For data shown in Extended Data Fig. 8d, RNA-Seq was performed with libraries prepared from the RNA samples collected from cells of the no deletion strain across 9 glucose concentrations and one *msn2* deletion strain at 0.2% glucose. RNA sample preparation was similar to the descriptions in the previous section. Library was constructed according to standard Illumina protocols. Sequencing was performed on a HiSeq 2500 sequencer. Both library construction and sequencing were performed at the core sequencing facility at Caltech. For the transient experiments, two biological replicates for each sample collected on different days were sequenced and analysed. Analysis of the sequencing data was performed with a local instance of Galaxy[55]. A standard analysis pipeline was used (alignment with Tophat[56]). Statistical test of differential expression between conditions was performed with duplicates using DESeq2[57].

Calculation of expected-by-chance fraction of overlapping pulsing. The heat map in Fig. 3g showed the expected fraction of Msn2 pulses that overlap with Mig1 pulses. This expected fraction measures the percentage of Msn2 pulses that would coincide with Mig1 pulses assuming the factors pulse independently of each other. Because an overlapping event is defined as when the peak of an Msn2 pulse falls into the time span of a Mig1 pulse, its expected fraction can be calculated as the fraction of time that Mig1 pulses occupy and is independent of Msn2 frequency, that is, $\frac{\text{number of Mig1 pulses per hour}\times\text{mean Mig1 duration}}{\text{1 hour}}$. As shown in Extended Data Fig. 8a, this calculated expected-by-chance overlap fraction is almost identical to the measured overlap fraction from an artificial population of cells where Msn2 and Mig1 dynamics are scrambled to enforce independence.

Fitting gene expression data with different models. In Extended Data Fig. 8b–d, we compared the ability of three models to fit combinatorial target gene expression levels across a range of glucose concentrations. The first model ('active–passive') includes both active and passive modulation, the second model ('passive only') includes only passive modulation, that is, assumes independent Msn2 and Mig1 dynamics, and the third model ('Msn2 only') assumes Mig1 does not reduce the effect of the Msn2 pulses (see Supplementary Discussion). In all models, gene expression is assumed to be activated by Msn2 and also occur at a basal level in the absence of nuclear Msn2. Mig1 is assumed to suppress both Msn2-activated (except in the Msn2 only model), and basal expression. In these models, expression is thus proportional to the frequency of effective Msn2 pulses (those not suppressed by Mig1 pulses, see definition below), plus the promoter-specific basal activity:

$$E^i_{\text{model}}=a\cdot f^i_{\text{Msn2eff}}+b\cdot\theta^i_{\text{Mig1out}}$$

Here, i labels the glucose condition; a denotes the mean amount of gene expression produced by each effective Msn2 pulse; f^i_{Msn2eff} is the frequency of effective Msn2 pulses per hour (calculated based on single-cell data, see details below); b is the basal promoter activity when Mig1 is out of the nucleus; and $\theta^i_{\text{Mig1out}}$ is the fraction of time that Mig1 is out of the nucleus (also calculated based on single-cell data). Note that the three models differ only in the effective Msn2 pulse frequency. In general, the active-passive model has the lowest f^i_{Msn2eff}, because

in this model Mig1 pulses suppress the effects of Msn2 pulses even more frequently than expected if Msn2 and Mig1 were independent, that is, in the 'passive only' model. In contrast, the 'Msn2 only' model has the highest f^i_{Msn2eff}.

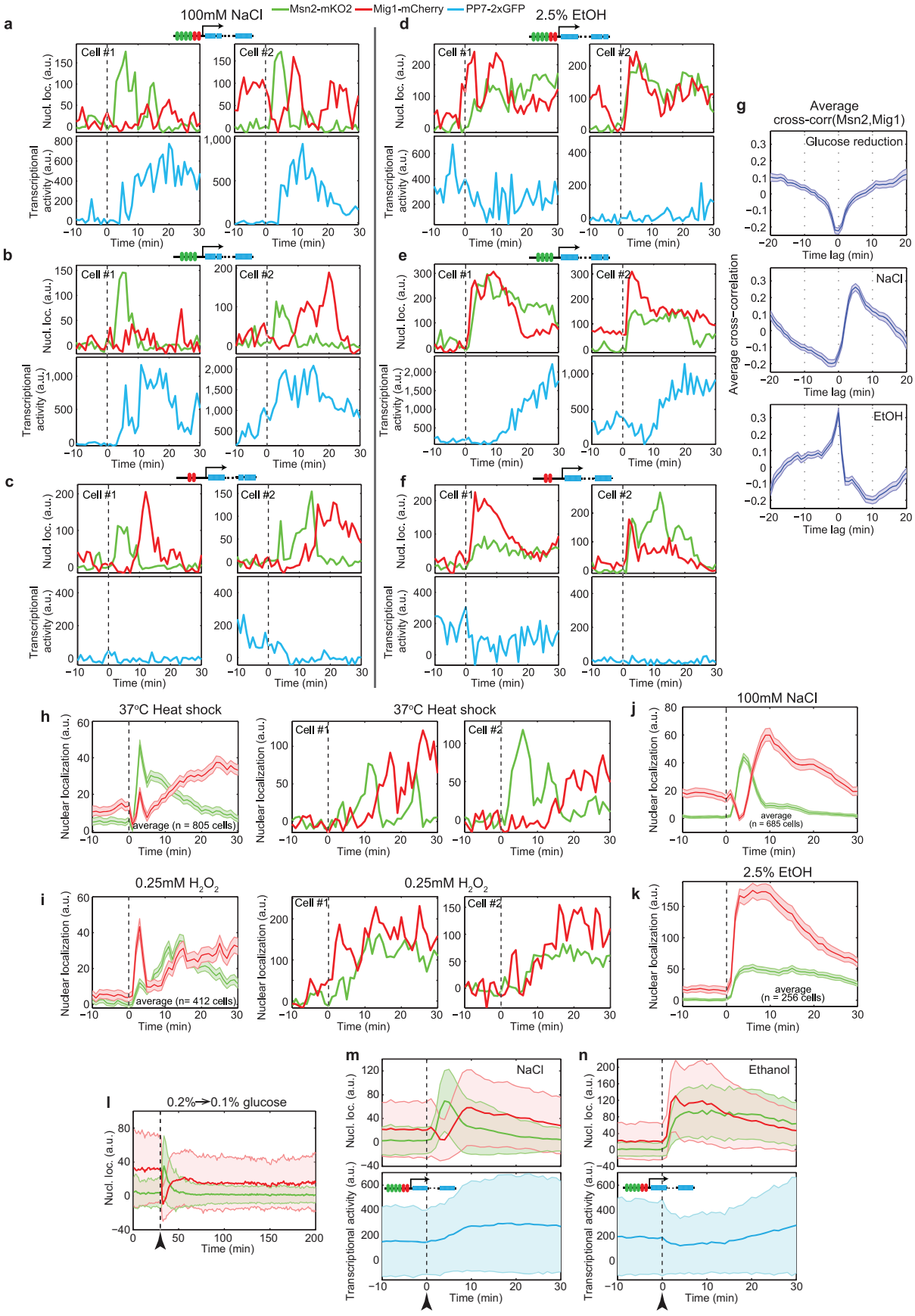
We calculated the effective Msn2 frequency, f^i_{Msn2eff}, with two different levels of temporal precision (see Supplementary Discussion). The simpler binary relative timing model considers Msn2 pulses to be either overlapping or non-overlapping with Mig1, as in Fig. 3. By contrast, the more precise continuous relative timing model allows for the empirically observed continuous dependence of expression level on the time interval between the Msn2 and Mig1 pulses, as shown in Extended Data Fig. 5b.

In the binary model, the effective Msn2 pulse frequency is simply the frequency of non-overlapping Msn2 pulses (Fig. 3). In the continuous model, the effect of an observed Msn2 pulse on a natural target's gene expression was determined by its pulse timing relative to Mig1 using the results in Extended Data Fig. 5b. More specifically, we normalized the data in Extended Data Fig. 5b such that Msn2 only pulses (those at the longest absolute time intervals) have a relative expression level of 1, while overlapping Msn2 pulses (time interval 0) have a relative expression level of 0. For each observed Msn2 pulse we calculated an effective gene expression contribution based on its timing relative to Mig1. This calculation was performed across all traces and all glucose concentrations to obtain f^i_{Msn2eff}. Prior to fitting, we converted the relative qPCR expression data to an absolute scale (equivalent to FPKM, fragments per kilobase of transcript per million mapped reads) using the RNA-seq data at 0.05% glucose as a reference (Extended Data Fig. 2f). We also used RNA-seq data from an *msn2* mutant to independently estimate parameter *b*. Thus, for each of the three models, only the parameter *a* needs to be fit. The least-squares fitting was performed by minimizing the error function $\sum_{i=1}^{9}\left[E^i_{\mathrm{model}}-E^i_{\mathrm{exp}}\right]^2$, where E^i_{exp} denotes the experimentally measured gene expression levels at glucose level *i* from qPCR and RNA-seq data sets.

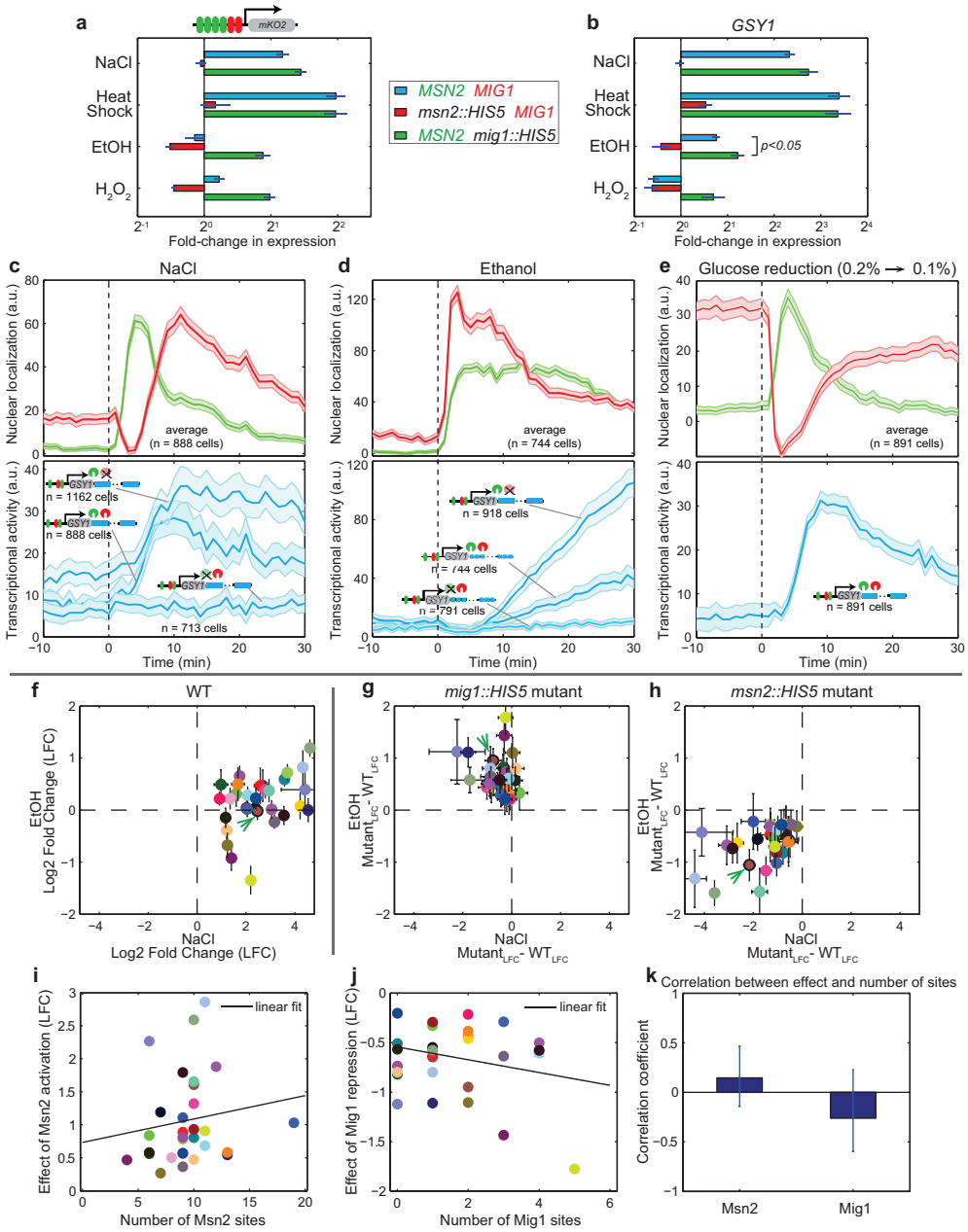
Statistical analysis. To compare single-cell data between different conditions, we computed the 95% confidence intervals of the sample mean for each set of single cells by the bootstrap method. More specifically, resampling with replacement was implemented with Matlab and 2,000 resamplings of the same sample size were obtained for each set of single cells. These 2,000 sets of single-cell data were then used for downstream analysis such as pulse-triggered averaging analysis and others. Bias-corrected 95% confidence interval[58] of the 2,000 samples were then calculated and represented as error bars or shaded regions. To compare distributions, the Kolmogorov–Smirnov test was used.

No statistical methods were used to predetermine sample size. The experiments were not randomized, and the investigators were not blinded to allocation during experiments and outcome assessment.

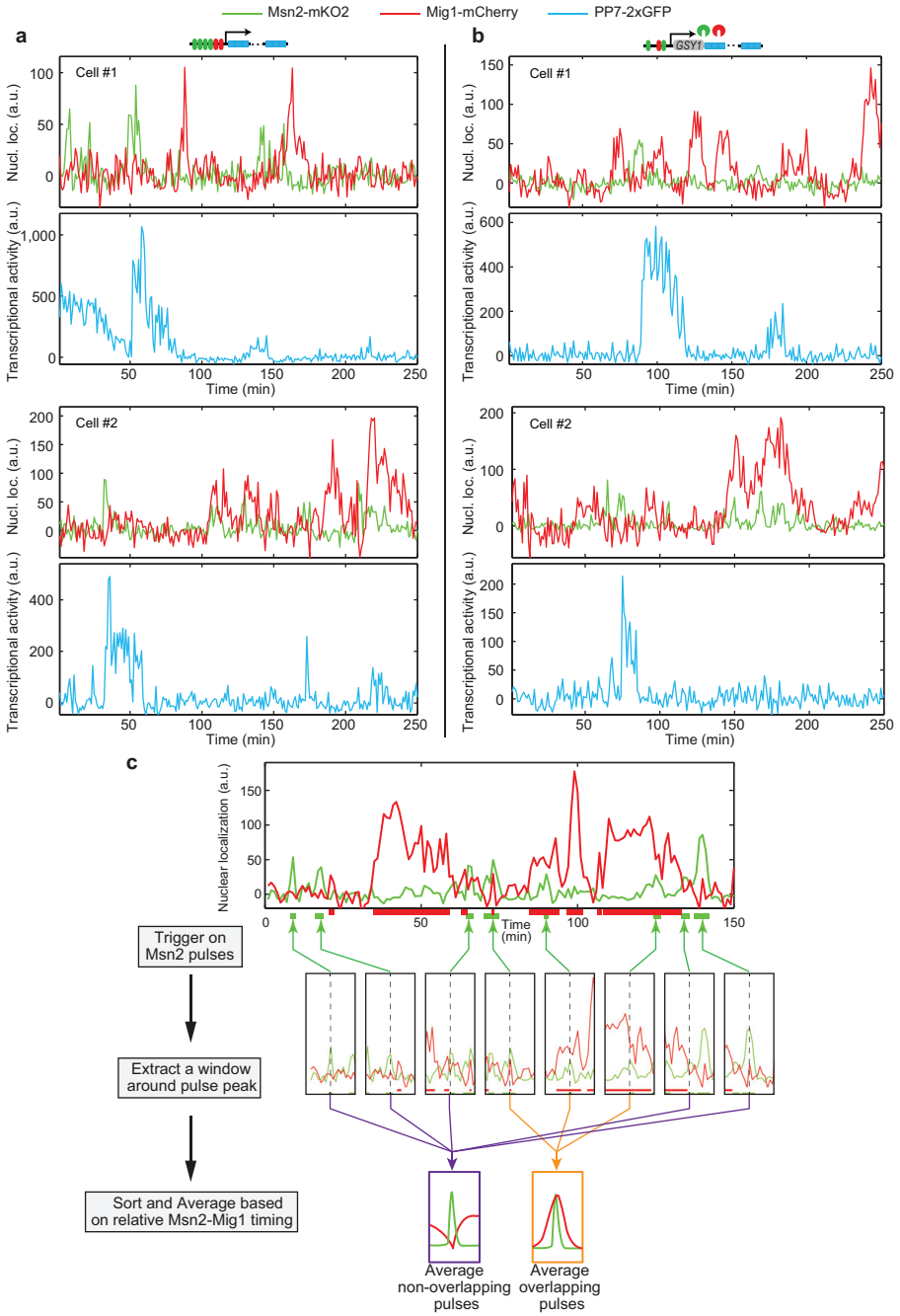
47. Gietz, R. D. & Schiestl, R. H. High-efficiency yeast transformation using the LiAc/SS carrier DNA/PEG method. *Nature Protocols* **2,** 31–34 (2007).
48. Sheff, M. A. & Thorn, K. S. Optimized cassettes for fluorescent protein tagging in *Saccharomyces cerevisiae*. *Yeast* **21,** 661–670 (2004).
49. Goldstein, A. L. & McCusker, J. H. Three new dominant drug resistance cassettes for gene disruption in *Saccharomyces cerevisiae*. *Yeast* **15,** 1541–1553 (1999).
50. Lutfiyya, L. L. *et al.* Characterization of three related glucose repressors and genes they regulate in *Saccharomyces cerevisiae*. *Genetics* **150,** 1377–1391 (1998).
51. Edelstein, A., Amodaj, N., Hoover, K., Vale, R. & Stuurman, N. Computer control of microscopes using μManager. *Curr. Protoc. Mol. Biol.* Unit 14.20 (2010).
52. Jaqaman, K. *et al.* Robust single-particle tracking in live-cell time-lapse sequences. *Nature Methods* **5,** 695–702 (2008).
53. Teste, M. A., Duquenne, M., François, J. M. & Parrou, J. L. Validation of reference genes for quantitative expression analysis by real-time RT-PCR in *Saccharomyces cerevisiae*. *BMC Mol. Biol.* **10,** 99 (2009).
54. Ye, J. *et al.* Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction. *BMC Bioinformatics* **13,** 134 (2012).
55. Goecks, J., Nekrutenko, A., Taylor, J., The Galaxy Team. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol.* **11,** R86 (2010).
56. Trapnell, C. *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature Protocols* **7,** 562–578 (2012).
57. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15,** 550 (2014).
58. Efron, B. *The Jackknife, the Bootstrap, and Other Resampling Plans* (Society for Industrial and Applied Mathematics, 1982).



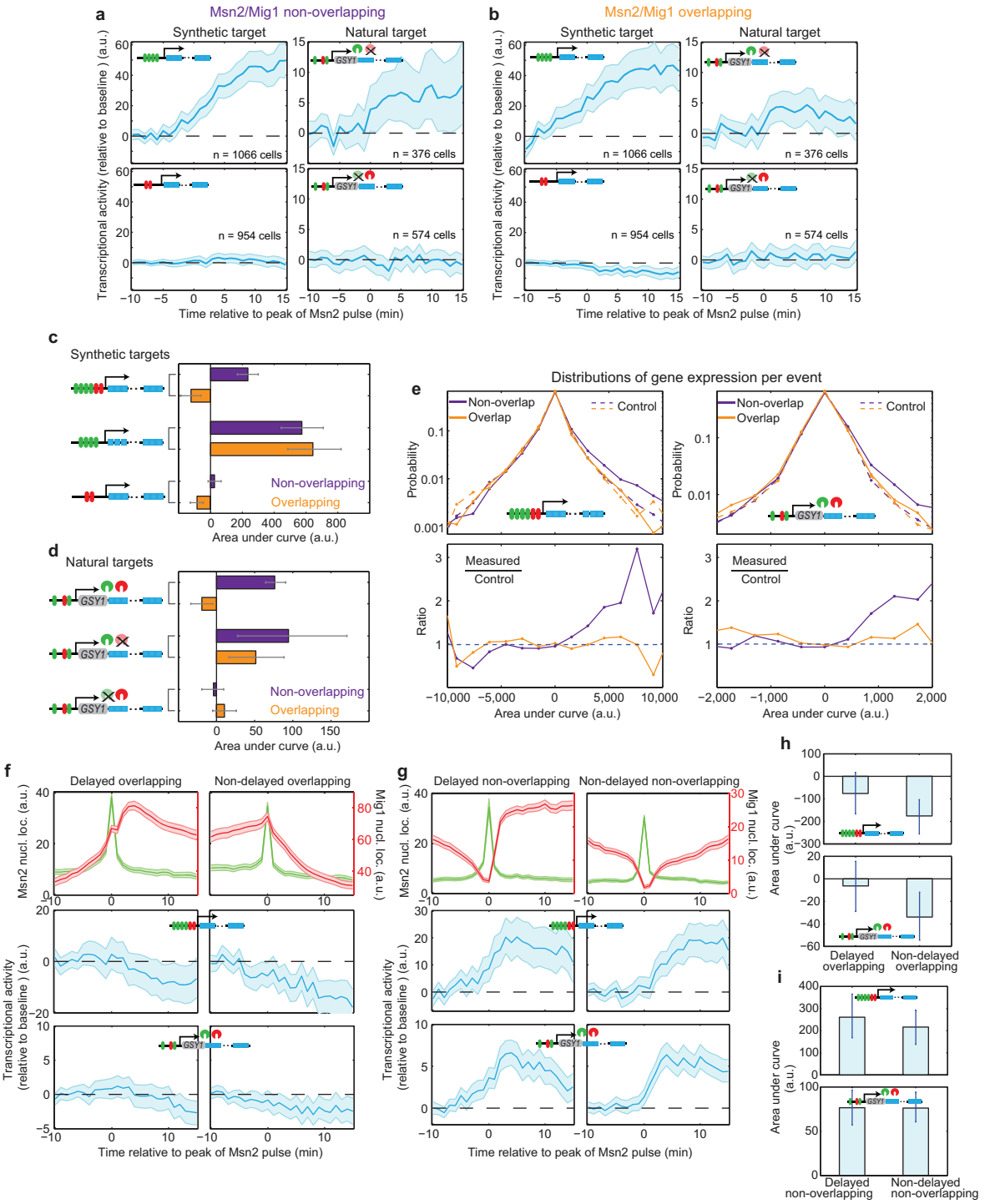
Extended Data Figure 1 | Single-cell analysis of relative pulse timing modulation by stress identity during transient response. **a**–**c**, Example traces for synthetic combinatorial (**a**), Msn2-specific (**b**), or Mig1-specific (**c**) promoters, in response to addition of 100 mM NaCl. Two cells are shown for each strain. For each cell, Msn2 and Mig1 localization traces (green and red) and the corresponding promoter response (blue) are shown on separate panels (top and bottom). Vertical dashed line indicates time of NaCl addition. **d**–**f**, Similar example traces for the response to addition of 2.5% ethanol. **g**, Average cross-correlation function of the transient Msn2 and Mig1 responses from $t = 0$–30 min after indicated stress. Cross-correlation between Msn2 and Mig1 is negative at time lag zero for both glucose reduction and NaCl stresses, but positive for ethanol stress. **h**, **i**, Averaged (left) and single-cell (right) nuclear localization traces of Msn2–mKO2 and Mig1–mCherry in response to 37 °C heat shock (**h**) or 0.25 mM H_2O_2 (**i**). **j**, **k**, Msn2 and Mig1 dynamics observed in Fig. 2b, c do not depend on the deletions introduced to the strain background. Averaged nuclear localization traces of Msn2–mKO2 and Mig1–mCherry in response to 100 mM NaCl (**j**) or 2.5% ethanol (**k**) for a control strain without *msn4 mig2* deletions. Shading indicates 95% confidence intervals of the mean. **l**–**n**, Standard deviation representations of different sets of single-cell data (presented in main figures). The mean is indicated with a solid line, and ± 1 standard deviation ranges are indicated by shading. **l**, Nuclear localization responses of Msn2–mKO2 (green) and Mig1–mCherry (red) to downshift in glucose level (see Fig. 1b). **m**, **n**, Nuclear localizations and transcriptional responses to NaCl and ethanol. (see Fig. 2b, c).



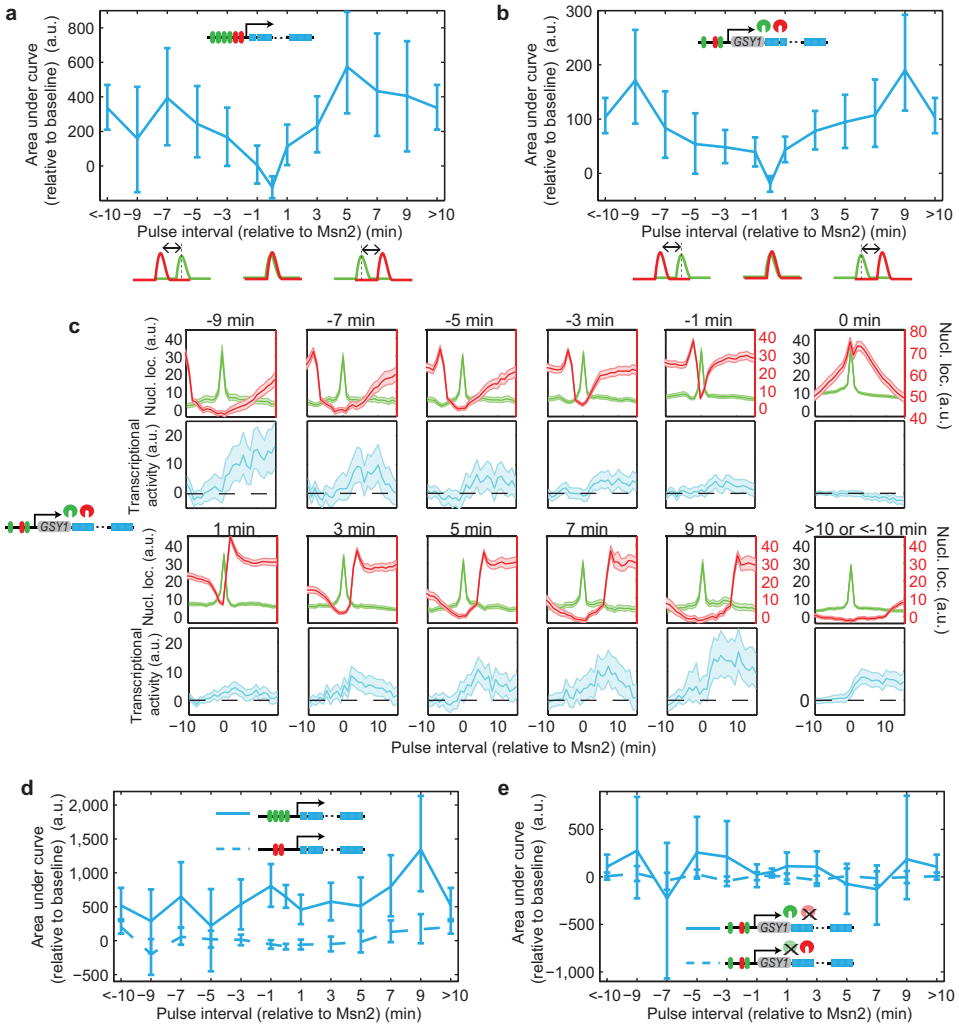
Extended Data Figure 2 | Additional data and analysis for transient stress responses. **a**, Fold-change in expression in response to different stresses for synthetic combinatorial target gene for three genetic backgrounds: no deletion (*MSN2 MIG1*, data from Fig. 2f), *msn2* deletion, and *mig1* deletion. **b**, Similar plot for the endogenous target gene *GSY1*. Cells were treated with designated stress for 10 min and ≥ 3 biological replicates were averaged (error bar indicates s.e.m.). *P* value was obtained from two-tailed *t*-test. **c**, **d**, Averaged transcriptional responses of *GSY1-24xPP7* in response to 100 mM NaCl (**c**) or 2.5% ethanol (**d**) for three genetic backgrounds: no deletion, *mig1* deletion, and *msn2* deletion. Averaged nuclear localization traces of Msn2–mKO2 and Mig1–mCherry for the 'no deletion' strain are shown on the top panels. **e**, Averaged nuclear localization traces of Msn2–mKO2 and Mig1–mCherry (top) and corresponding transcriptional responses for *GSY1-24xPP7* in response to glucose downshift (from 0.2% to 0.1%). Shading in **c**–**e** indicates 95% confidence intervals of the mean. **f**–**k**, RNA-seq analysis (see Methods and Supplementary Discussion for more details). **f**, $\log_2$ fold-changes (LFC) in gene expression of 31 identified combinatorial targets (including *GSY1*; brown circle, indicated by green arrow) in response to NaCl (*x* axis) and ethanol (*y* axis) for wild-type background (that is, no deletion of either *MSN2* or *MIG1*). **g**, The differences in LFC between wild-type and *mig1* deletion for both NaCl (*x* axis) and ethanol (*y* axis). **h**, The differences in LFC between wild-type and *msn2* deletion for both NaCl (*x* axis) and ethanol (*y* axis). **i**, The effect of Msn2 for each target was plotted against the corresponding number of Msn2 binding sites. **j**, Analogous plot for the effect of Mig1 binding sites. **k**, Correlation coefficients between the effect of Msn2 or Mig1 and the number of Msn2 or Mig1 binding motif, respectively. Error bars in **f**–**h** indicate standard deviations from two biological replicates. Error bars in **k** represent 95% confidence intervals from bootstrap.



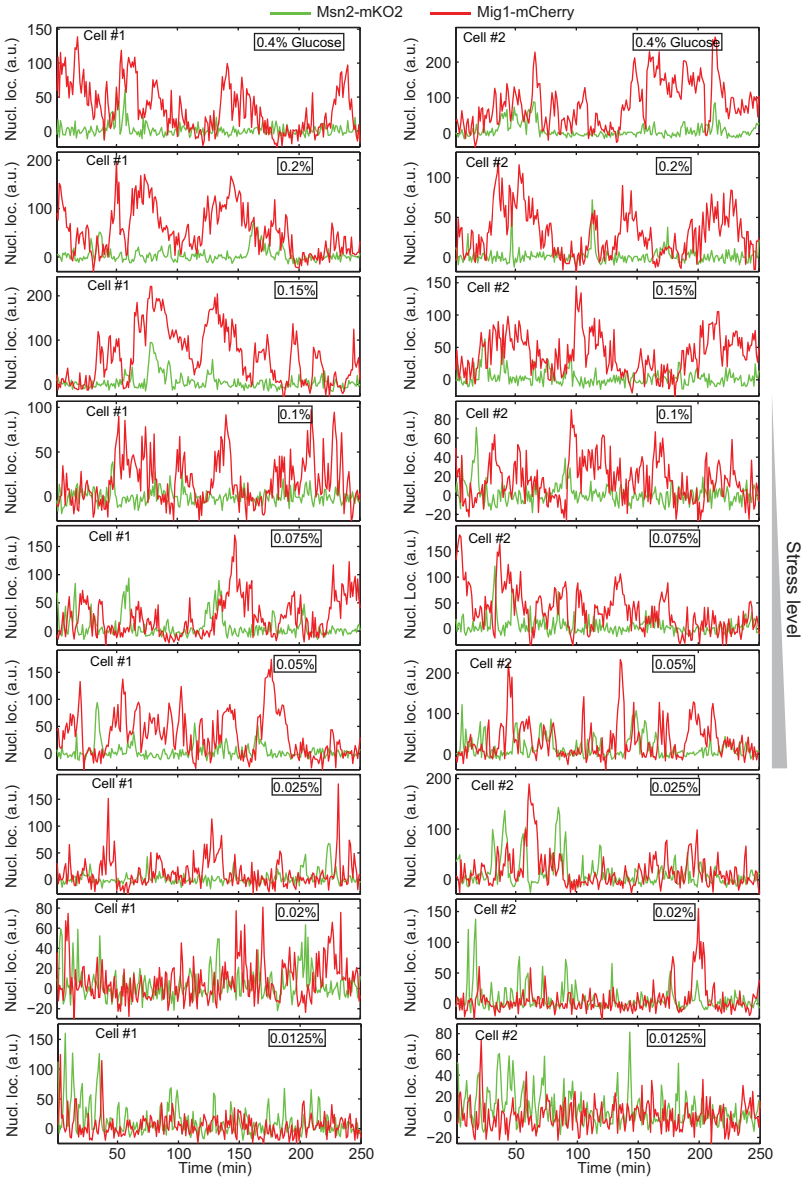
Extended Data Figure 3 | Example 3-colour single-cell traces under steady-state conditions, and schematic diagram of pulse-triggered averaging analysis. **a**, **b**, Example 3-colour single-cell traces for synthetic (**a**) and natural (**b**) promoters under constant glucose (0.05%). Two cells are shown for each promoter. For each cell, nuclear localization traces are shown on the top and PP7–2 × GFP transcriptional output signal is shown on the bottom. **c**, Schematic illustration of pulse-triggered averaging analysis. Msn2 pulses were identified (green arrows) and sorted based on their relationship with the Mig1 signal within a 21 min time window (see Methods). Horizontal green and red lines underneath top time trace plot indicate width of identified Msn2 and Mig1 pulses, respectively. Msn2 pulses whose peaks overlap with Mig1 pulses were categorized as overlapping events (orange arrows) while the rest of Msn2 pulses were categorized as non-overlapping events (purple arrows). Overlapping and non-overlapping events were then averaged separately (bottom schematics).



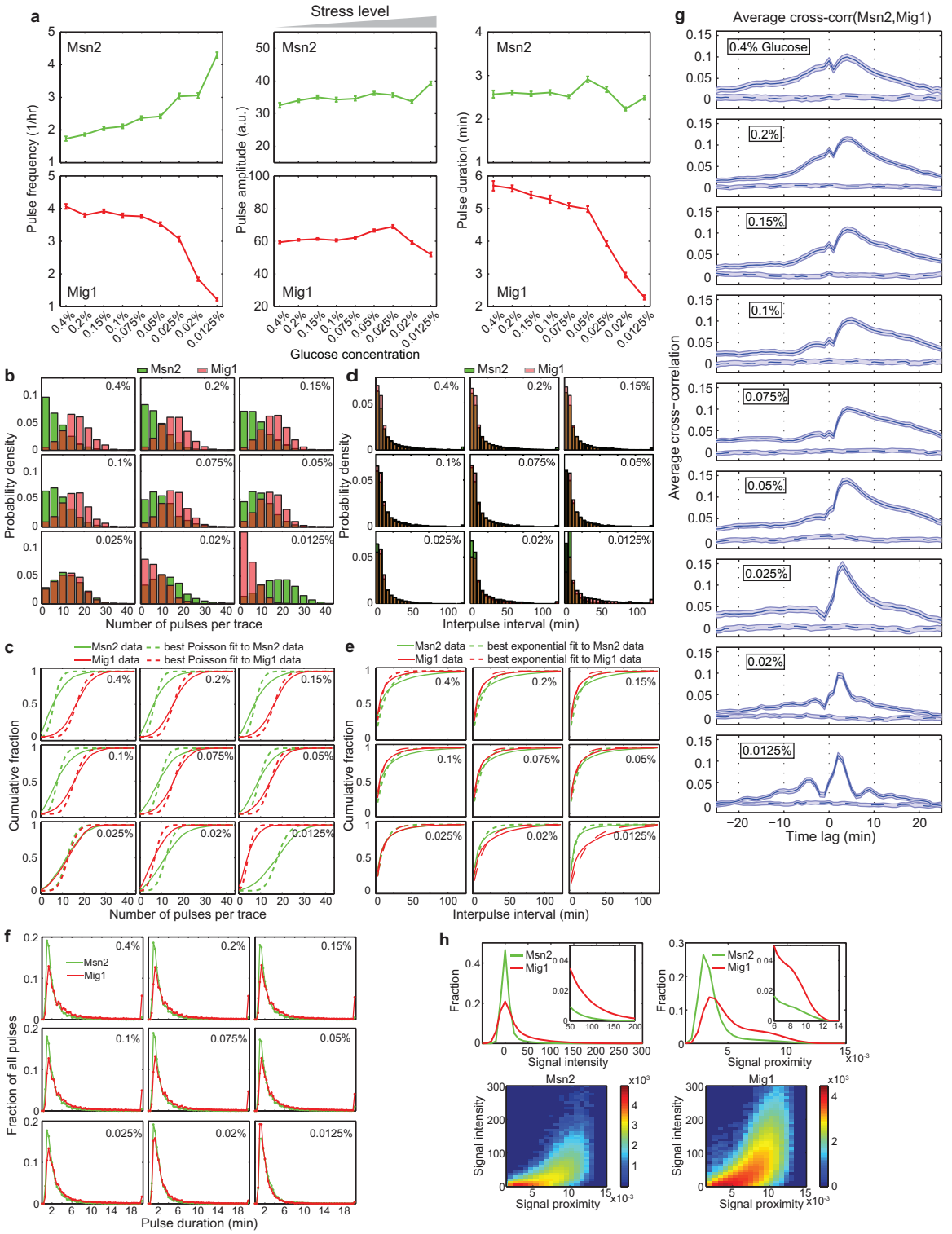
Extended Data Figure 4 | Pulse-triggered averaging analysis for control promoters and for delayed pulse timing events. **a**, **b**, Plots analogous to those in Fig. 3d, e for additional synthetic and natural promoters. The *GSY1* promoter was examined in strains with Msn2 or Mig1 zinc-finger deletions. For gene expression, areas under curves were analysed and presented in **c**, **d**. **c**, Relative pulse timing-dependent gene expression occurs for combinatorial promoters but not pure Msn2 or Mig1 target promoters. Bars represent integrated gene expression based on area under curve from Fig. 3d, e and **a**, **b**. **d**, Plot analogous to **c** for the natural *GSY1* target gene. Binding of the transcription factors was abolished by mutations in zinc finger DNA-binding domains, indicated by crosses. **e**, Distributions of gene expression (estimated as integrated area under curve) per non-overlapping or overlapping event for both synthetic and natural combinatorial promoters (real data (solid) versus control data (dashed); top) and ratios between real and control data (bottom). Control data was measured from scrambled population of cells. For the real data, the distributions of non-overlapping and overlapping events are significantly different (by Kolmogorov–Smirnov test) with P values of 2.1×10^{-17} and 1.2×10^{-15} for synthetic and natural promoters, respectively. In contrast, for control data, they are not significantly different (P values: 0.4520 and 0.9888). For the calculation of ratios, averages of the non-overlapping and overlapping control data were used as control. **f**–**i**, Pulse-triggered averaging analysis of 'delayed' events in which an Msn2 pulse is followed by a Mig1 pulse (see Supplementary Discussion for details). **f**, Overlapping events were subdivided into delayed and non-delayed depending, as shown. Corresponding mean Msn2 and Mig1 signals as well as transcriptional responses were plotted for both synthetic and natural promoters. A similar classification was performed for non-overlapping events (**g**). Area under curve for **f**, **g** was plotted for direct comparison of gene expression between delayed and non-delayed pulse timing events (**h**, **i**). Shading and error bars indicate 95% confidence intervals of the mean. Schematic promoters indicate whether the synthetic or natural *GSY1* promoter were used in each case.



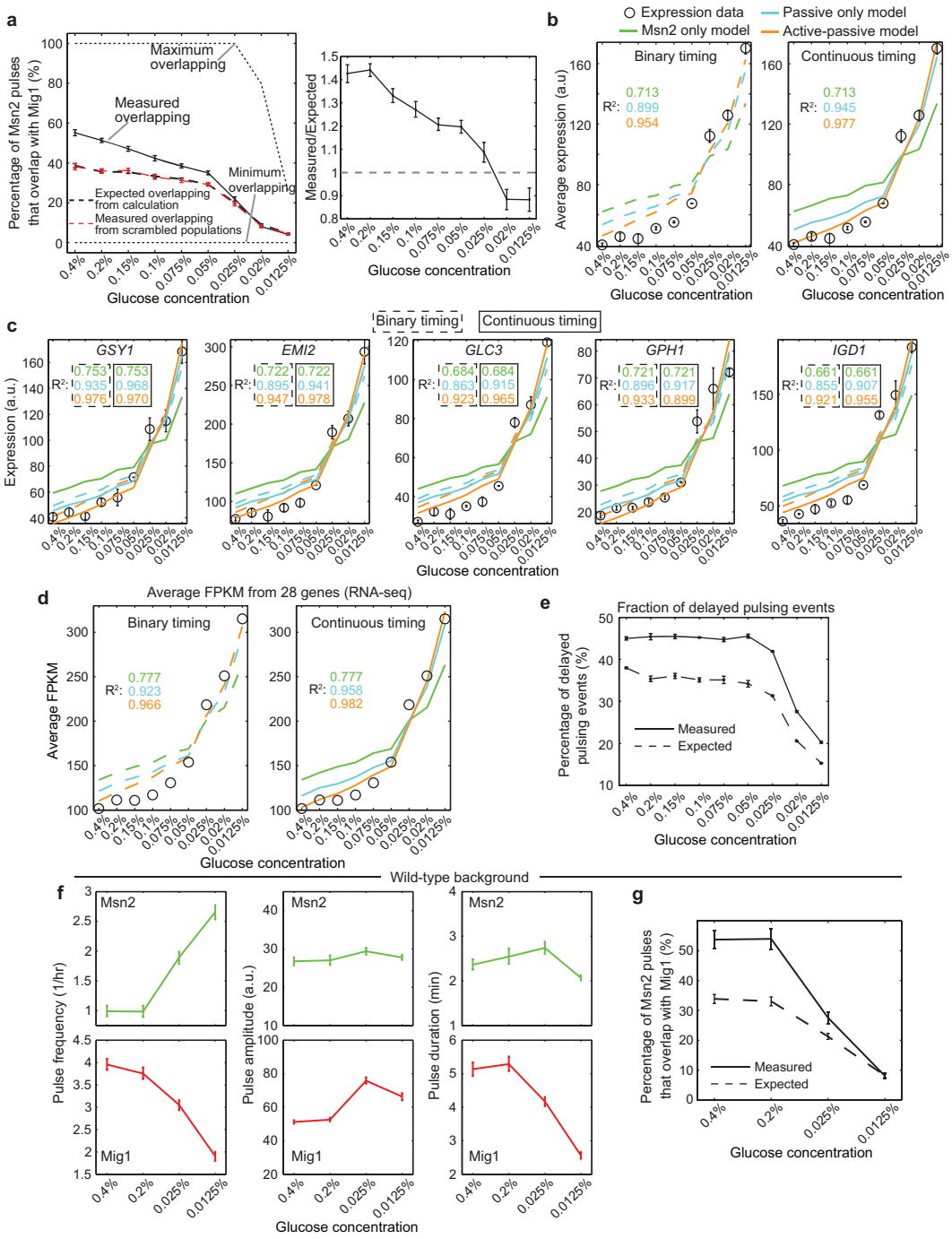
Extended Data Figure 5 | Analysis of mean gene expression dependence on time interval (continuous relative timing) between Msn2 and Mig1 pulses. **a**, **b**, Mean expression from both synthetic (**a**) and natural (**b**) target promoters depends on the time interval between Msn2 and Mig1 pulses (that is, interval between the peak of an Msn2 pulse and the edge of the nearest Mig1 pulse). For each time interval, mean expression values were determined by integrating the area under the baseline-subtracted averaged PP7 traces, and averaging within bins of similar pulse interval. **c**, Specifically, Msn2 pulses were categorized on the basis of the pulse interval between Msn2 and Mig1 and the corresponding PP7 signals were averaged and their areas under curve were plotted (Methods). The pulse interval ranges from −9 to 9 min, which represents the bin centre of each 2-min bin (for example, 1 min represents the range 2 min ≥ interval > 0 min), with the 0 min interval representing overlapping events. Both >10 or <−10 min intervals represent events where Msn2 pulses were not surrounded by any Mig1 pulses within 21 min. **d**, **e**, Msn2 and Mig1 regulation are both necessary for continuous relative timing-dependent gene expression under constant glucose condition. Analysis similar to **a**, **b** was performed on synthetic Msn2- and Mig1-specific promoters (**d**) and natural *GSY1* promoter with Msn2 or Mig1 zinc finger deletion mutants (**e**). Shading and error bars indicate 95% confidence intervals of the mean.



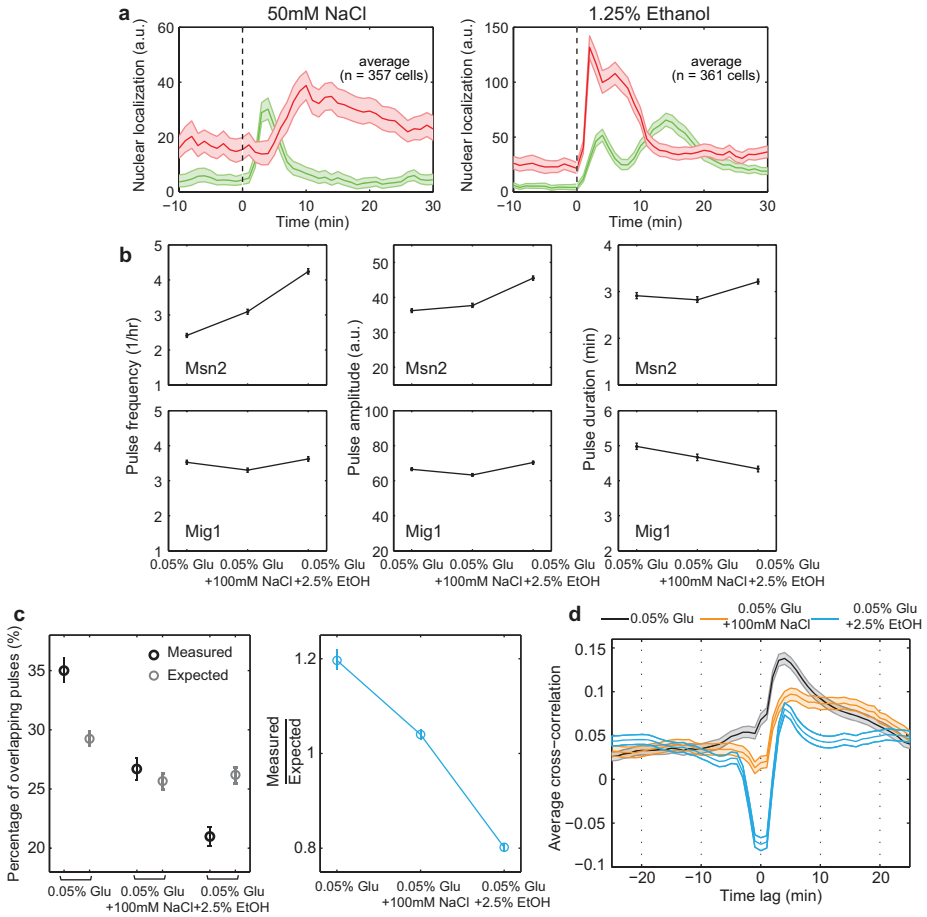
Extended Data Figure 6 | Example single-cell nuclear localization traces for different constant glucose conditions. Two single-cell traces are shown for each indicated glucose level (boxed percentage values). Cells were switched to indicated glucose level from 0.2% glucose at 110 min before time zero (that is, beginning of movie acquisition).



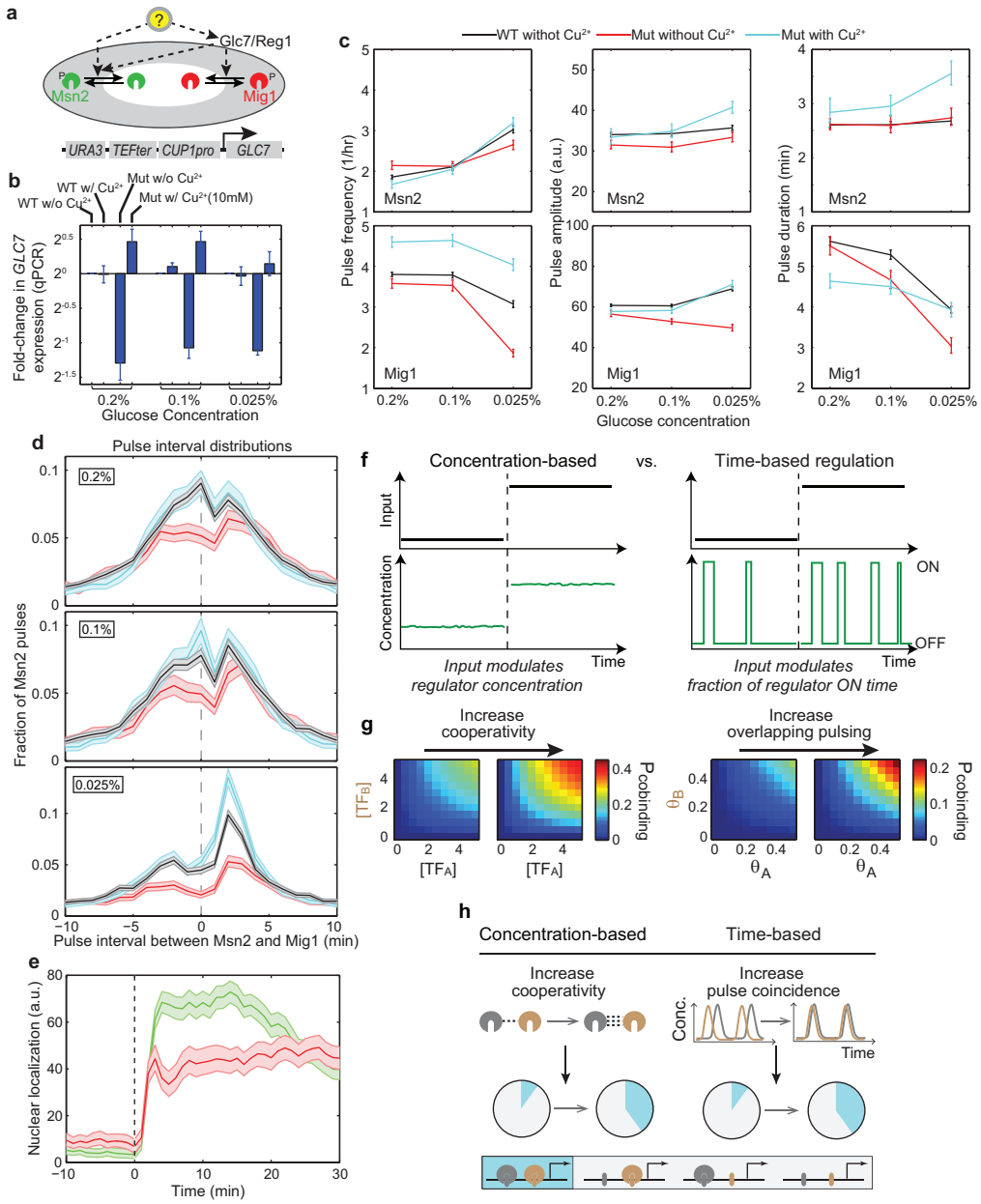
Extended Data Figure 7 | Characterization of Msn2 and Mig1 pulses and average cross-correlation functions between Msn2 and Mig1 in individual cells across different constant glucose concentrations. **a**, Pulse frequency, amplitude, and duration analysis. Single-cell traces at each glucose level were analysed and the mean frequency, amplitude and duration for both Msn2 and Mig1 were plotted. **b**, **c**, Distributions of total number of pulses per trace across glucose concentrations (**b**), along with corresponding fits to Poisson distributions (shown as cumulative distributions, **c**). Kolmogorov–Smirnov (KS) tests showed that these distributions differ significantly from Poisson distributions ($P < 10^{-16}$). **d**, **e**, Analogous plots for the distributions of inter-pulse time intervals (**d**), and corresponding fits to exponential distributions (**e**). These distributions differ significantly from exponential distributions according to KS tests ($P < 10^{-57}$). **f**, Distributions of pulse duration for Msn2 and Mig1 across glucose concentrations. **g**, Cross-correlation function (solid blue) of Msn2 and Mig1 nuclear localization traces, that is, cross-corr(Msn2, Mig1) (Methods). Dashed blue lines represent negative (independent) controls, calculated by scrambling the Msn2-Mig1 trace pairs within a population of cells (that is, cross-correlating Msn2 from one cell with Mig1 from another, randomly chosen, cell). Shading and error bars indicate 95% confidence intervals of the mean. The number of cells analysed in each glucose concentration: 1,511 (0.4%), 3,475 (0.2%), 2,605 (0.15%), 2,075 (0.1%), 3,034 (0.075%), 2,768 (0.05%), 1,392 (0.025%), 2,055 (0.02%), and 1,906 (0.0125%). **h**, Two different localization metrics show similar Msn2 and Mig1 state distributions. Top left, histogram of the intensity score for Msn2 and Mig1 shows long-tailed distributions for both proteins with peaks around zero (basal state). Insert, zoomed-in view of the tails. Top right, analogous plots for the signal proximity score also show long-tailed distributions with clear basal states. Signal proximity is the inverse of the distance-based localization metric described in the Methods section. High signal proximity indicates that the top 10 brightest pixels in the cell are close to each other. (Bottom) Signal intensity positively correlates with signal proximity for both Msn2 and Mig1, suggesting that these two independent scores show related features. This data are for cells at 0.05% glucose. Similar behaviours are observed across other glucose concentrations.



Extended Data Figure 8 | Further characterization of relative pulse timing modulation under steady-state conditions. **a**, Left, experimentally measured overlapping fraction (solid black) can be compared to minimum and maximum possible overlapping fractions (bottom and top dashed lines, respectively). The expected overlapping fraction for independent Msn2 and Mig1 dynamics is determined two ways: either computed from the Mig1 duty cycle (dashed black), or measured from scrambled populations (dashed red). Minimum and maximum possible fractions were calculated with the measured duty cycles of Msn2 and Mig1 pulses. Right, the ratios of measured overlapping fraction to expected overlapping fraction across glucose concentrations. **b**, Relative pulse timing modulation explains gene-expression dependence on glucose level for combinatorial target promoters. Black circles represent mean expression of 5 genes measured by qPCR (see Methods for normalization). Data were fit with three models, as indicated. See Methods and Supplementary Discussion for more details on binary and continuous timing models. R^2 values for fits are indicated in corresponding colours. Error bars indicate s.e.m. calculated from 3 biological replicates. **c**, Expression data for the 5 individual genes fit to the binary timing (dashed lines; R^2 values in dashed box) as well as continuous timing (solid lines; R^2 values in solid box) models. **d**, Analysis of RNA-seq expression data across 9 glucose concentrations. The averaged expression levels from 28 of the 31 identified combinatorial targets (Extended Data Fig. 2f–k) were fit with the binary or continuous timing modulation models (left and right plots, respectively). Three genes were excluded because they did not display a monotonic dependence on glucose (*YER067C-A*, *YKR098C*, *YLR109W*). In this analysis, parameter b was independently estimated from an *msn2* mutant at 0.2% glucose (samples collected on the same day). **e**, Glucose level modulates the fraction of delayed pulse timing events (see also Extended Data Fig. 4 and Supplementary Discussion). Total fractions of delayed overlapping (see Extended Data Fig. 4e, left) and delayed non-overlapping pulse events (see Extended Data Fig. 4f, left) were plotted across glucose concentrations. Expected fractions were computed from 'scrambled' populations where Msn2 and Mig1 dynamics are, by construction, independent. **f**, **g**, Glucose concentration also modulates relative pulse timing in a control strain without deletions of *msn4* and *mig2*. **f**, Pulse characteristics of both Msn2 and Mig1 for varying glucose concentrations. **g**, Measured versus expected overlapping fractions across different glucose concentrations (see **a**) for the wild-type background that was not deleted for Msn4 and Mig2. Error bars indicate 95% confidence intervals of the mean (except for **b**–**d**). The number of cells analysed for **f**, **g**: 618 (0.4%), 541 (0.2%), 714 (0.025%), and 775 (0.0125%).



Extended Data Figure 9 | Additional effects of stress level and type on transient and steady-state responses. **a**, Stress level does not modulate relative pulse timing during transient responses. Averaged nuclear localization traces of Msn2–mKO2 and Mig1–mCherry during transient response to 50 mM NaCl (left) or 1.25% ethanol (right) are shown (see Fig. 2b, c). **b**, Additional stresses modulate relative timing during steady-state responses. Changes in pulse characteristics of both Msn2 and Mig1 in response to the addition of 100 mM NaCl or 2.5% ethanol during steady-state growth at 0.05% glucose. **c**, Measured (black) versus expected (grey) overlapping fractions for the same 3 conditions as in **b**. **d**, Averaged cross-correlation between Msn2 and Mig1 time traces for the same three conditions. See Supplementary Discussion for additional discussion. Shading and error bars indicate 95% confidence intervals of the mean. The number of cells analysed for **b**, **d**: 2,768 (0.05% glucose), 2,178 (0.05% glucose with 100 mM NaCl) and 2,115 (0.05% glucose with 2.5% ethanol).



Extended Data Figure 10 | A role for Glc7 in active relative pulse timing modulation under constant glucose conditions and functional aspect of relative pulse timing modulation. **a**, Schematic of potential mechanisms for Glc7-dependent relative pulse timing modulation (top) and construct design (bottom). Overlapping pulsing of Msn2 and Mig1 could be induced by either a common kinase/phosphatase (such as Glc7) that directly or indirectly activates both Msn2 and Mig1 localization, or by an upstream input (yellow circle) that simultaneously regulates kinases/phosphatases responsible for Msn2 and Mig1 localization. To analyse the role of *GLC7* in relative pulse timing, we constructed a strain in which the normal *GLC7* promoter is replaced by a copper-inducible promoter, as shown. **b**, qPCR characterization of the inducible *GLC7* strain across three glucose concentrations. Basal copper level in the media reduced *GLC7* expression to less than 50% of its wild-type level. Addition of 10 μM $CuSO_4$ restored the expression to 110% to 140% of wild-type level. **c**, Changes in pulse characteristics in response to *GLC7* reduction (red) and restoration (blue), compared to wild-type (black). **d**, Corresponding changes in pulse interval distribution. Pulse interval was calculated as the distance between the peak of a given Msn2 pulse and the peak of its closest Mig1 pulse within a 21 min window. **e**, Averaged nuclear localization traces of Msn2–mKO2 (green) and Mig1–mCherry (red) in response to 2.5% ethanol addition (dashed line) for the *GLC7* reduction mutant. See Supplementary Discussion for additional discussion. Error bars in **b** indicate s.e.m. from 3 biological replicates. For **c**–**e**, shading and error bars indicate 95% confidence intervals of the mean. The number of cells analysed in the mutant strain: 671 (0.2% glucose without Cu^{2+}), 540 (0.1% glucose without Cu^{2+}), 719 (0.025% glucose without Cu^{2+}), 756 (0.2% glucose with Cu^{2+}), 643 (0.1% glucose with Cu^{2+}), and 656 (0.025% glucose with Cu^{2+}). **f**–**h**, Functional aspect of relative pulse timing modulation (see Supplementary Note). **f**, Concentration-based versus time-based regulation. Input modulates the regulator concentration (left) versus the fraction of regulator ON time (right). **g**, Modulation of relative pulse timing in time-based regulation results in changes in the effective protein–protein cooperativity. Increasing protein–protein cooperativity in concentration-based regulation changes the probability of co-binding of TF_A and TF_B (left). Increasing overlapping pulsing in time-based regulation leads to qualitatively similar changes in the probability of co-binding (right). Protein cooperativity parameter ω_{AB} was increased from 1 to 2 for the left plots. Overlap fraction was increased from $\theta_A\theta_B$ to $2 \times \theta_A\theta_B$ for the right plots ($\omega_{AB} = 1$). $K_A = K_B = 5$ for both left and right. **h**, Schematic, relative pulse timing modulation affects the relative probability of simultaneous binding of two transcription factors to a target promoter (right). This effect is analogous to that generated by cooperative protein–protein interactions (left)[44]. Stronger protein–protein interactions or a higher overlap fraction can both increase the probability with which two transcription factors will be simultaneously bound at neighbouring sites (schematic pie charts).

ARTICLE

doi:10.1038/nature15709

Ion channels enable electrical communication in bacterial communities

Arthur Prindle[1], Jintao Liu[1]*, Munehiro Asally[2]*, San Ly[1], Jordi Garcia-Ojalvo[3] & Gürol M. Süel[1]

The study of bacterial ion channels has provided fundamental insights into the structural basis of neuronal signalling; however, the native role of ion channels in bacteria has remained elusive. Here we show that ion channels conduct long-range electrical signals within bacterial biofilm communities through spatially propagating waves of potassium. These waves result from a positive feedback loop, in which a metabolic trigger induces release of intracellular potassium, which in turn depolarizes neighbouring cells. Propagating through the biofilm, this wave of depolarization coordinates metabolic states among cells in the interior and periphery of the biofilm. Deletion of the potassium channel abolishes this response. As predicted by a mathematical model, we further show that spatial propagation can be hindered by specific genetic perturbations to potassium channel gating. Together, these results demonstrate a function for ion channels in bacterial biofilms, and provide a prokaryotic paradigm for active, long-range electrical signalling in cellular communities.

Communication through electrical signalling is prevalent among biological systems, with one of the most familiar examples being the action potential in neurons that is mediated by ion channels[1]. For many years, the study of bacterial ion channels has provided fundamental insights into the structural basis of such neuronal signalling[2,3]. In particular, the prokaryotic potassium ion channel KcsA provided the first structural information on ion selectivity and conductance[4]. More recently, it has been shown that bacteria possess many important classes of other ion channels, such as sodium channels[5], chloride channels[6], calcium-gated potassium channels[7] and ionotropic glutamate receptors[8], similar to those found in neurons. However, the native role of these ion channels in bacteria has largely remained unclear[9,10]. Efforts to uncover ion channel function in bacteria have identified roles in the extreme acid resistance response[6] and in osmoregulation[11], yet ion-specific channels do not appear to be solely responsible for these cellular processes. It remains unclear whether ion channels can support other unique functions in prokaryotes. We hypothesized that studying bacteria in their native context, the biofilm community, may reveal new clues about the function of ion channels in bacteria.

Bacterial biofilms are organized communities containing billions of densely packed cells. Such communities can exhibit fascinating macroscopic spatial coordination[12–17]. However, it remains unclear how microscopic bacteria can communicate effectively over large distances. To investigate this question, we studied a *Bacillus subtilis* microbial community that was recently reported to undergo metabolic oscillations triggered by nutrient limitation[18]. The oscillatory dynamics resulted from long-range metabolic co-dependence between cells in the interior and periphery of the biofilm (Fig. 1a)[18]. Specifically, interior and peripheral cells compete for glutamate, while sharing ammonium. As a result, biofilm growth halts periodically, increasing nutrient availability for the sheltered interior cells. Interestingly, glutamate (Glu^-) and ammonium (NH_4^+) are both charged metabolites, whose respective uptake and retention is known to depend on the transmembrane electrical potential and proton motive force[19,20]. Therefore, we wondered whether metabolic coordination among distant cells within the biofilm might also involve a form of electrochemical signalling.

Oscillations in membrane potential

To monitor long-range electrical fluctuations in the bacterial community as a function of space and time, we grew biofilms in an unconventionally large microfluidic device (Fig. 1b and 'Microfluidics' section of Methods). To measure electrical signalling, we used the fluorescent cationic dye thioflavin T (ThT) to quantify membrane potential within the biofilm. ThT is positively charged and can be retained in cells because of the negative electrical membrane potential inside cells. Thus, cells with a negative membrane potential will retain more ThT, allowing it to act as a Nernstian voltage indicator[21,22]. We confirmed that ThT faithfully reports the membrane potential by comparing it to an established reporter of membrane potential in bacteria[23], 3,3′-dipropylthiadicarbocyanine iodide ($DiSC_3(5)$) (Extended Data Fig. 1a). We found that ThT has an approximately threefold higher sensitivity to changes in membrane potential compared to $DiSC_3(5)$ (Extended Data Fig. 1a, inset). Furthermore, we exposed cells to minor changes in external pH, which is known to alter membrane potential[24], and observed the expected changes in ThT (Extended Data Fig. 1b). Therefore, ThT accurately reports on changes in membrane potential for bacteria residing in biofilms.

We next investigated changes in membrane potential during metabolic oscillations. In particular, quantitative measurements of ThT fluorescence showed global and self-sustained oscillations consistent with the reported period of the metabolic oscillations (Fig. 1c, Supplementary Video 1 and Extended Data Fig. 1c)[18]. Furthermore, oscillations in ThT could be quenched by supplementation of the media with glutamine, which bypasses the need for glutamate and ammonium (Extended Data Fig. 1d). These data show a connection between metabolic oscillations and membrane potential. Notably, oscillations in membrane potential were synchronized among even the most distant regions of the biofilm community (Fig. 1d, e). We wondered whether active electrochemical signalling could be responsible for this long-range synchronization.

[1]Division of Biological Sciences, University of California San Diego, California 92093, USA. [2]Warwick Integrative Synthetic Biology Centre, School of Life Sciences, University of Warwick, Coventry CV4 7AL, UK. [3]Department of Experimental and Health Sciences, Universitat Pompeu Fabra, 08003 Barcelona, Spain.
*These authors contributed equally to this work.

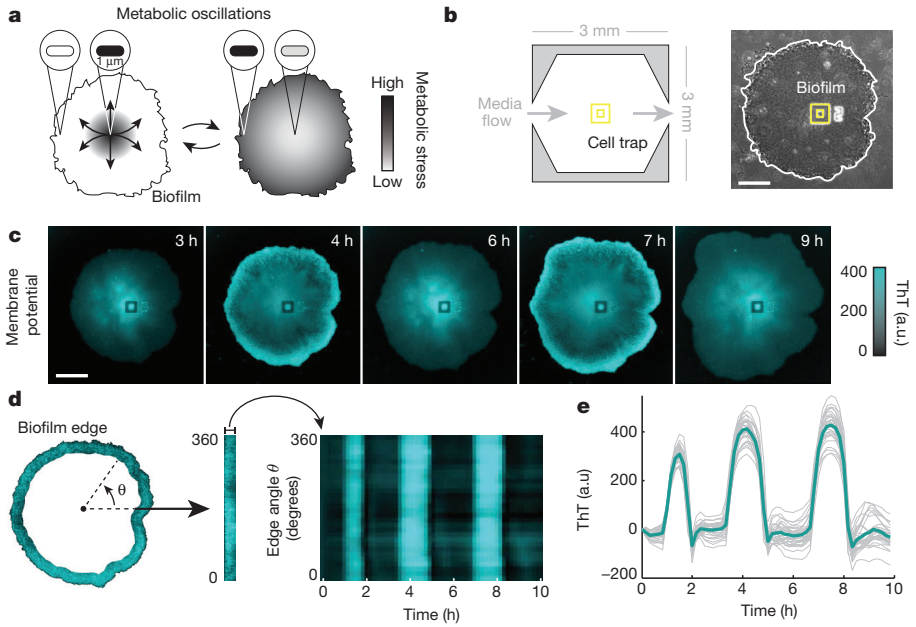


Figure 1 | Biofilms produce synchronized oscillations in membrane potential. **a**, Biofilms generate collective metabolic oscillations resulting from long-range metabolic interactions between interior and peripheral cells[18]. It remains unclear how microscopic bacteria are capable of communicating over such macroscopic distances within biofilm communities. **b**, Schematic of the microfluidic device used throughout this study (left). Phase contrast image of a biofilm growing in the microfluidic device with the cell trap highlighted in yellow (right). Scale bar, 100 μm. **c**, Global oscillations in membrane potential, as reported by thioflavin T (ThT), within the biofilm community. ThT is positively charged but not known to be actively transported, so it can be retained in cells due to their negative membrane potential inside the cell. ThT fluorescence increases when the inside of the cell becomes more negative, and thus ThT is inversely related to the membrane potential. Scale bar, 0.15 mm. Representative images shown are taken from over 75 independent biofilms. a.u., arbitrary units. **d**, Membrane potential oscillations are highly synchronized even between the most distant regions of the biofilm. To analyse synchronization, the edge region of the biofilm was identified and straightened (left) then plotted over time (right). **e**, Time traces of the heat map shown in **d**. Indicated in bold is the mean of 30 traces.

Active propagation of potassium signal

Changes in membrane potential involve the movement of charged species across the cellular membrane. We suspected the involvement of potassium, since it is the most abundant cation in all living cells[25] and has been implicated to have a role in biofilm formation[26,27]. *B. subtilis* uses active potassium transport mechanisms to concentrate intracellular potassium at approximately 300 mM[28–30]. This intracellular concentration is nearly 40 times the external media concentration. Consequently, sudden release of this potassium gradient would increase extracellular potassium concentration and generate a change in the membrane potential. Accordingly, we used a fluorescent chemical potassium dye, asante potassium green-4 (APG-4[31]), to measure the extracellular concentration of potassium in the biofilm (Fig. 2a and Extended Data Fig. 2a, b). We observed global oscillations in APG-4 that correlated with membrane potential, which suggests that the membrane potential oscillations could involve the release of potassium (Fig. 2b, c and Supplementary Video 2). In agreement with this finding, oscillations in extracellular potassium extended beyond the biofilm to the surrounding growth media (Extended Data Fig. 2c). We also measured the dynamics of sodium, another ion commonly used by cells to modulate membrane potential, and observed no oscillations (Extended Data Fig. 2d–f). Together, these data suggest that potassium has a role in the synchronized oscillations in membrane potential.

Furthermore, we directly tested that oscillations in membrane potential were driven by flow of potassium across the cell membrane. Specifically, we clamped net potassium flux across the cell membrane by supplementing the growth media with 300 mM KCl (matching the intracellular potassium concentration) (Fig. 2d). When we applied this chemical potassium clamp, oscillations in membrane potential abruptly halted (Fig. 2e). Applying this clamp together with valinomycin, a potassium ionophore that acts as potassium-specific carrier in the cellular membrane[32], yielded a similar quenching of oscillations (Extended Data Fig. 3a, b). Therefore, changes in the electrochemical potential for potassium appear to be required for the observed oscillations in membrane potential.

Next, we determined whether cells could actively propagate the extracellular potassium signal through the biofilm to sustain long-range communication. While diffusive signals decay over space and time, active signalling processes can amplify the signal, avoiding such decay (Fig. 2f). To determine which of these processes may be operating in the biofilm, we observed the propagation of the extracellular potassium signal (Fig. 2g). Results show that the signal travels at a constant rate of propagation (Extended Data Fig. 3c, d). Furthermore, the amplitude of the signal does not decay with distance travelled, in contrast to what is predicted for passive potassium diffusion (Fig. 2h). These findings are consistent with a process in which cells actively propagate the potassium signal. Together, these results suggest that the biofilm synchronizes global oscillations in membrane potential by an active signalling process involving potassium ions.

Potassium ion-channel-mediated signalling

Motivated by our findings, we explored the role of ion channels in the observed potassium signalling. We focused on YugO, the only experimentally described potassium channel in *B. subtilis*, which is also reported to be important for biofilm formation[33]. Potassium flux through YugO is gated by an intracellular TrkA domain, known to be regulated by the metabolic state of the cell[34–36]. Accordingly, we hypothesized that metabolic limitation could form the initial trigger for YugO activation. Specifically, since glutamate limitation is known to drive the underlying metabolic oscillations[18], we anticipated that transient removal of glutamate could initiate potassium release. To test this, we transiently deprived cells of glutamate and measured extracellular potassium in both wild-type and *yugO* deletion strains (see 'Strains' section of the Methods). As expected, we observed extracellular potassium increase for wild-type but not the *yugO* deletion

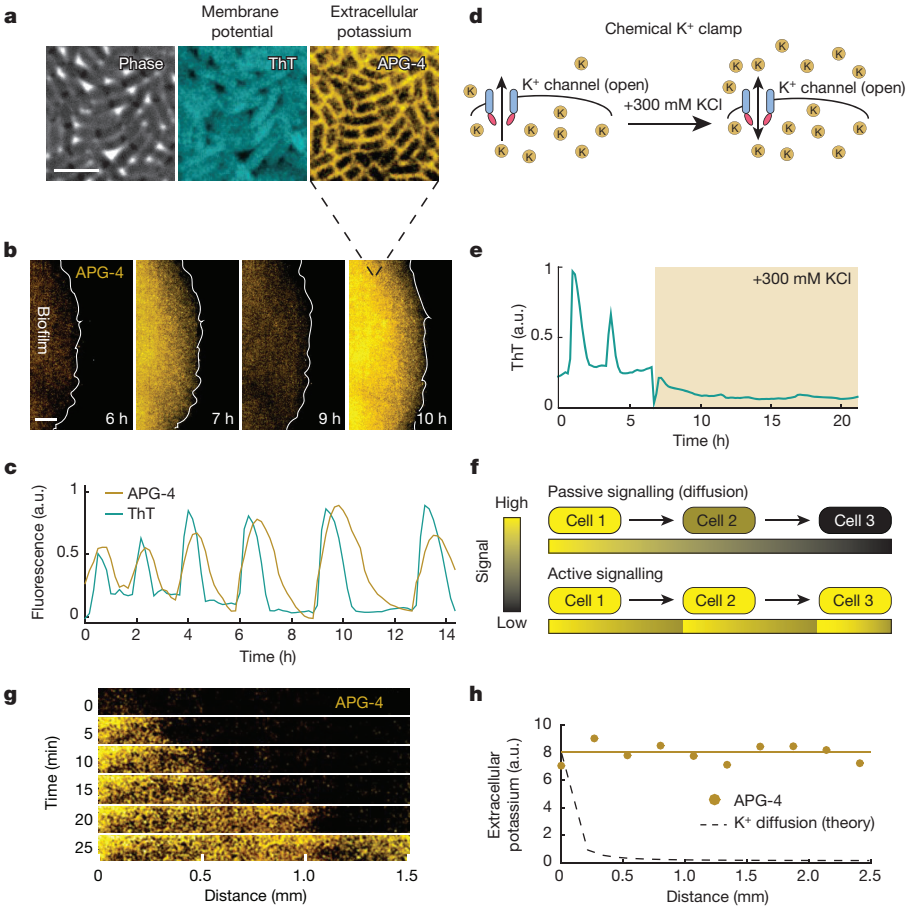


Figure 2 | Potassium release is involved in active signal propagation within the biofilm. **a**, An extracellular fluorescent chemical dye (APG-4) reports the concentration of potassium in the media (Extended Data Fig. 2a, b). For comparison, the same cells are shown stained with ThT, which is inversely related to the membrane potential. These images depict cells at the peak of the ThT oscillation cycle. Representative images are selected from six independent experiments. Scale bar, 2 μm. **b**, Global oscillations in extracellular potassium throughout the biofilm. A white line indicates the edge of the biofilm. Representative images are selected from six independent experiments. Scale bar, 0.2 mm. **c**, Oscillations in membrane potential and extracellular potassium are synchronized, suggesting that potassium release is involved in global membrane potential oscillations. ThT is inversely related to the membrane potential. Representative traces are taken from the experiment shown in **b**. **d**, A chemical potassium clamp (300 mM KCl, matching the intracellular concentration[29]) prevents the formation of potassium electrochemical gradients across the cellular membrane. **e**, Clamping net potassium flux quenches oscillations in membrane potential. Representative trace is selected from two independent experiments. **f**, Illustration of the differences between passive signalling (diffusion) and active signalling. When cells passively respond to a signal, the range that the signal can propagate is limited due to the decay of signal amplitude. In contrast, when cells actively respond by amplifying the signal, propagation can extend over greater distances. **g**, We measured propagation of extracellular potassium by measuring APG-4 in time and along a length of approximately 1.5 mm within the biofilm. **h**, Extracellular potassium amplitude is relatively constant as the signal propagates, in contrast to the predicted amplitude decay of a passive signal. Representative data selected from six independent experiments. The diffusion line is calculated using the 2D diffusion equation and the diffusion coefficient for potassium within biofilms (Supplementary Information).

strain (Fig. 3a). These findings suggest that glutamate limitation can trigger the potassium signal via the YugO potassium channel.

Next, we investigated whether YugO also has a role in the active propagation of the potassium signal. To test this, we measured the response of wild-type and *yugO* deletion strains to transient bursts of external potassium (300 mM KCl). As expected, potassium exposure first resulted in a short-term membrane potential depolarization in both strains. However, in the wild-type strain this initial depolarization was typically followed by an extended hyperpolarization phase, which was not observed in the *yugO* deletion strain (Fig. 3b). This period of hyperpolarization was accompanied by an increase in extracellular potassium (Extended Data Fig. 4a). Together, these data indicate that potassium exposure triggers a release of intracellular potassium through YugO. Exposure to an equivalent concentration of sorbitol (an uncharged solute) did not elicit an equivalent response, ruling out purely osmotic effects (Extended Data Fig. 4b). Therefore, YugO appears to have a role in propagating the extracellular potassium signal within the biofilm.

Mathematical modelling of electrical signalling

Our data thus point to a proposed mechanism where metabolically stressed cells release intracellular potassium, and the resulting elevated extracellular potassium imposes further metabolic stress onto neighbouring cells (Fig. 3c). In *B. subtilis*, glutamate is co-transported with two protons by the GltP transporter and this process depends on the proton motive force[19]. Potassium-mediated depolarization of the membrane potential can transiently reduce the electrical component of the proton motive force[24], and thereby lower glutamate uptake and intracellular ammonium retention[19,20]. Therefore, potassium-mediated signalling could propagate metabolic stress onto distant cells (Fig. 3c, right). Accordingly, hyperpolarization triggered by YugO activation may represent a cellular response to enhance glutamate uptake or ammonium retention. This notion is supported by our finding that the response to extracellular potassium can be abolished by growing cells in glutamine, an uncharged metabolite and preferred nitrogen source that bypasses the need for glutamate and ammonium[37] (Extended Data Fig. 4c). This result further supports the

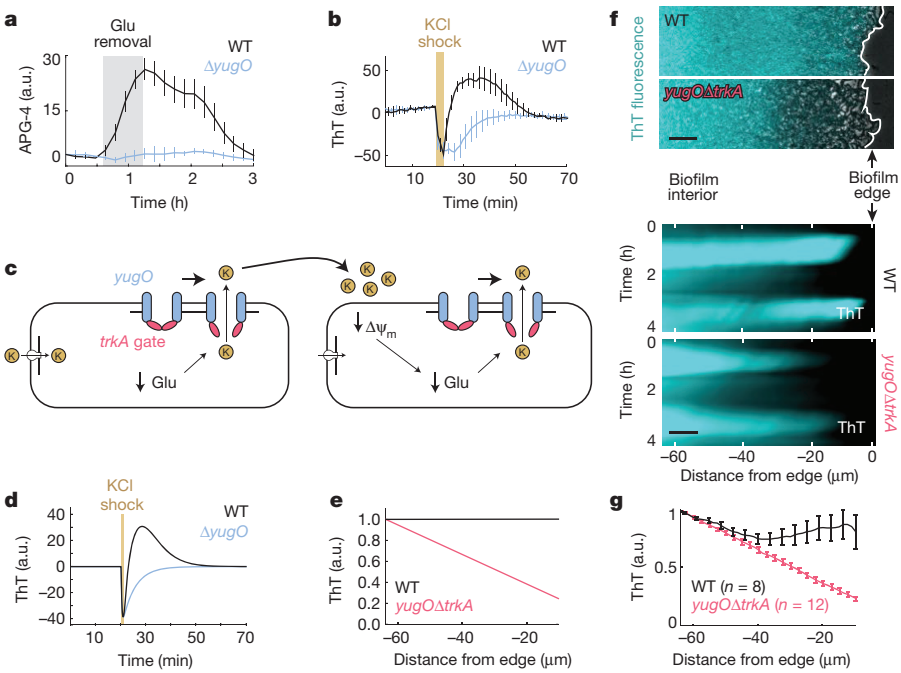


Figure 3 | The molecular mechanism of signal propagation involves potassium channel gating. **a**, *yugO* is a potassium channel in *B. subtilis* that is gated intracellularly by a *trkA* domain, which is regulated by the metabolic state of the cell[34–36]. Withdrawing glutamate (the sole nitrogen source in MSgg media) induces an increase in extracellular potassium (APG-4) for wild-type (WT) but not the *yugO* deletion strain. Error bars indicate the mean ± s.d. for three independent biofilms each. **b**, An external potassium shock (300 mM KCl) induces a short-term membrane potential depolarization in both wild-type and *yugO* deletion strains. However, in the wild type this initial depolarization was followed by hyperpolarization, which is not observed in the *yugO* deletion strain (mean ± s.d. for 12 traces drawn from 3 biofilms each). ThT is inversely related to the membrane potential. **c**, Proposed model for potassium signalling. The initial trigger for potassium release is metabolic stress caused by glutamate limitation. External potassium depolarizes neighbouring cells, producing further nitrogen limitation by limiting glutamate uptake, and thus produces further metabolic stress. This cycle results in cell–cell propagation of the potassium signal. **d**, A minimal conductance-based model describing the dynamics of the cell's membrane potential in terms of a single potassium channel and a leak current. Consistent with our experimental results, this simple model exhibits transient depolarization followed by hyperpolarization in response to local increases in extracellular potassium concentration. **e**, The model predicts that manipulating channel gating and conductance will result in decaying amplitude in the spatial propagation of membrane potential oscillations. **f**, Maximum intensity projection of membrane potential change illustrating attenuated communication within the biofilm in a *yugOΔtrkA* deletion compared to wild-type biofilms (top). Heat map of oscillations taken from wild-type and *yugOΔtrkA* mutant biofilms (bottom). Representative images are taken from three independent biofilm experiments in which wild-type and *yugOΔtrkA* biofilms are compared head-to-head. Scale bars, 8 μm. **g**, Quantification of normalized pulse amplitude from wild-type ($n = 8$ pulses) and *yugOΔtrkA* ($n = 12$ pulses) mutant biofilms (mean ± s.e.m.).

specific link between potassium-mediated electrical signalling and metabolic stress.

To determine whether the proposed potassium-channel-based mechanism is sufficient to account for the observed propagating pulses of electrical activity, we turned to mathematical modelling. Specifically, we considered a minimal conductance-based model describing the dynamics of the cell's membrane potential in terms of a single potassium channel and a leak current (see 'Mathematical Model' section of the Supplementary Information). Consistent with our experimental results, this simple model exhibits transient depolarization followed by hyperpolarization in response to local increases in extracellular potassium concentration (Fig. 3d). Furthermore, the model shows long-range propagation of these excitations without decay in the amplitude of membrane potential oscillations (Fig. 3e). Therefore, the proposed mechanism is mathematically sufficient to qualitatively account for the observed membrane potential dynamics and active propagation in space.

The model also predicts that reduced efficiency of the potassium channel function could lead to degradation in long-range communication (Fig. 3e). Since a complete *yugO* deletion interferes with development of large biofilms[33], we constructed a strain in which we deleted the TrkA gating domain, leaving only the ion channel portion of YugO intact (see 'Strains' section of the Methods). Similarly truncated bacterial potassium channels have been shown to have altered gating and ion conductance[34,38]. Indeed, the *yugOΔtrkA* mutant biofilms exhibited a reduced propagation of membrane potential oscillations (Fig. 3f and Supplementary Video 3). Specifically, in contrast to wild type, the *yugOΔtrkA* mutant shows decay in the signal amplitude from the interior of the biofilm to the cells at the periphery, which is also consistent with model predictions (Fig. 3g). Thus, YugO channel gating appears to promote efficient electrical communication between distant cells.

Discussion

Our findings suggest that bacteria use potassium ion-channel-mediated electrical signals to coordinate metabolism within the biofilm. The ensuing 'bucket brigade' of potassium release allows cells to rapidly communicate their metabolic state, taking advantage of a link between membrane potential and metabolic activity. This form of electrical communication can thus enhance the previously described long-range metabolic co-dependence in biofilms[18]. Specifically, the wave of depolarization triggered by metabolically stressed interior cells would limit the ability of cells in the biofilm periphery to take up glutamate or retain ammonium, thereby allowing interior cells more access to these nutrients. This also provides a possible explanation for the observation that the *yugO* deletion strain has a defect in biofilm development[33]. Interestingly, owing to the rapid diffusivity of potassium ions in aqueous environments, it is also conceivable that even physically disconnected biofilms could be capable of synchronizing their metabolic oscillations by a similar exchange of potassium ions.

The role of ion-channel-mediated electrical communication has long been appreciated[39]. While cation channels are found in all organisms[9,10]

and potassium is the dominant intracellular cation[25], electrical signalling is commonly viewed to be a property of neurons. However, several recent studies have suggested that in addition to traditional cell-to-cell communication systems such as quorum sensing[40], bacteria may use electron flux[41–43] to communicate. The herein described study of electrical coordination of metabolism in microbial communities may in turn hold some general insights that extend beyond bacteria. For example, the connection between neuronal signalling and metabolic activity (neurometabolism) is an active area of research[44,45]. Furthermore, depletion of glutamate, the most common excitatory neurotransmitter[46], also forms the initial trigger for these collective metabolic oscillations synchronized by potassium. Therefore, it is intriguing to think not only about the structural similarities between bacterial and human potassium ion channels[2,3], but also their possible functional similarities with respect to long-range electrical communication.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 16 June; accepted 10 September 2015.

Published online 21 October 2015.

1. Gerstner, W. & Kistler, W. M. *Spiking Neuron Models: Single Neurons, Populations, Plasticity* (Cambridge Univ. Press, 2002).
2. Hille, B. *Ion Channels of Excitable Membranes* (Sinauer Associates, 2001).
3. MacKinnon, R. Potassium channels and the atomic basis of selective ion conduction. *Biosci. Rep.* **24,** 75–100 (2004).
4. Doyle, D. A. *et al.* The structure of the potassium channel: molecular basis of K^+ conduction and selectivity. *Science* **280,** 69–77 (1998).
5. Ren, D. *et al.* A prokaryotic voltage-gated sodium channel. *Science* **294,** 2372–2375 (2001).
6. Iyer, R., Iverson, T. M., Accardi, A. & Miller, C. A biological role for prokaryotic ClC chloride channels. *Nature* **419,** 715–718 (2002).
7. Jiang, Y. *et al.* Crystal structure and mechanism of a calcium-gated potassium channel. *Nature* **417,** 515–522 (2002).
8. Chen, G. Q., Cui, C., Mayer, M. L. & Gouaux, E. Functional characterization of a potassium-selective prokaryotic glutamate receptor. *Nature* **402,** 817–821 (1999).
9. Kuo, M. M. C., Haynes, W. J., Loukin, S. H., Kung, C. & Saimi, Y. Prokaryotic K^+ channels: From crystal structures to diversity. *FEMS Microbiol. Rev.* **29,** 961–985 (2005).
10. Saimi, Y., Loukin, S. H., Zhou, X. L., Martinac, B. & Kung, C. Ion channels in microbes. *Methods Enzymol.* **294,** 507–524 (1998).
11. Martinac, B., Buechner, M., Delcour, A. H., Adler, J. & Kung, C. Pressure-sensitive ion channel in *Escherichia coli*. *Proc. Natl Acad. Sci. USA* **84,** 2297–2301 (1987).
12. Costerton, J. W., Stewart, P. S. & Greenberg, E. P. Bacterial biofilms: a common cause of persistent infections. *Science* **284,** 1318–1322 (1999).
13. Hall-Stoodley, L., Costerton, J. W. & Stoodley, P. Bacterial biofilms: from the natural environment to infectious diseases. *Nature Rev. Microbiol.* **2,** 95–108 (2004).
14. Vlamakis, H., Aguilar, C., Losick, R. & Kolter, R. Control of cell fate by the formation of an architecturally complex bacterial community. *Genes Dev.* **22,** 945–953 (2008).
15. Asally, M. *et al.* Localized cell death focuses mechanical forces during 3D patterning in a biofilm. *Proc. Natl Acad. Sci. USA* **109,** 18891–18896 (2012).
16. Wilking, J. N. *et al.* Liquid transport facilitated by channels in *Bacillus subtilis* biofilms. *Proc. Natl Acad. Sci. USA* **110,** 848–852 (2013).
17. Payne, S. *et al.* Temporal control of self-organized pattern formation without morphogen gradients in bacteria. *Mol. Syst. Biol.* **9,** 697 (2013).
18. Liu, J. *et al.* Metabolic co-dependence gives rise to collective oscillations within microbial communities. *Nature* **523,** 550–554 (2015).
19. Tolner, B., Ubbink-Kok, T., Poolman, B. & Konings, W. N. Characterization of the proton/glutamate symport protein of *Bacillus subtilis* and its functional expression in *Escherichia coli*. *J. Bacteriol.* **177,** 2863–2869 (1995).
20. Boogerd, F. C. *et al.* AmtB-mediated NH_3 transport in prokaryotes must be active and as a consequence regulation of transport by GlnK is mandatory to limit futile cycling of NH_4^+/NH_3. *FEBS Lett.* **585,** 23–28 (2011).
21. Kralj, J. M., Hochbaum, D. R., Douglass, A. D. & Cohen, A. E. Electrical spiking in *Escherichia coli* probed with a fluorescent voltage-indicating protein. *Science* **333,** 345–348 (2011).
22. Lo, C.-J., Leake, M. C., Pilizota, T. & Berry, R. M. Nonequivalence of membrane voltage and ion-gradient as driving forces for the bacterial flagellar motor at low load. *Biophys. J.* **93,** 294–302 (2007).
23. Strahl, H. & Hamoen, L. W. Membrane potential is important for bacterial cell division. *Proc. Natl Acad. Sci. USA* **107,** 12281–12286 (2010).
24. Krulwich, T. A., Sachs, G. & Padan, E. Molecular aspects of bacterial pH sensing and homeostasis. *Nature Rev. Microbiol.* **9,** 330–343 (2011).
25. Epstein, W. The roles and regulation of potassium in bacteria. *Prog. Nucleic Acid Res. Mol. Biol.* **75,** 293–320 (2003).
26. López, D., Fischbach, M. A., Chu, F., Losick, R. & Kolter, R. Structurally diverse natural products that cause potassium leakage trigger multicellularity in *Bacillus subtilis*. *Proc. Natl Acad. Sci. USA* **106,** 280–285 (2009).
27. Kinsinger, R. F., Kearns, D. B., Hale, M. & Fall, R. Genetic requirements for potassium ion-dependent colony spreading in *Bacillus subtilis*. *J. Bacteriol.* **187,** 8462–8469 (2005).
28. Vieira-Pires, R. S., Szollosi, A. & Morais-Cabral, J. H. The structure of the KtrAB potassium transporter. *Nature* **496,** 323–328 (2013).
29. Holtmann, G., Bakker, E. P., Uozumi, N. & Bremer, E. KtrAB and KtrCD: Two K^+ uptake systems in *Bacillus subtilis* and their role in adaptation to hypertonicity. *J. Bacteriol.* **185,** 1289–1298 (2003).
30. Whatmore, A. M., Chudek, J. A. & Reed, R. H. The effects of osmotic upshock on the intracellular solute pools of *Bacillus subtilis*. *J. Gen. Microbiol.* **136,** 2527–2535 (1990).
31. Rimmele, T. S. & Chatton, J. Y. A novel optical intracellular imaging approach for potassium dynamics in astrocytes. *PLoS ONE* **9,** 1–9 (2014).
32. Margolin, Y. & Eisenbach, M. Voltage clamp effects on bacterial chemotaxis. *J. Bacteriol.* **159,** 605–610 (1984).
33. Lundberg, M. E., Becker, E. C. & Choe, S. MstX and a putative potassium channel facilitate biofilm formation in *Bacillus subtilis*. *PLoS ONE* **8,** e60993 (2013).
34. Cao, Y. *et al.* Gating of the TrkH ion channel by its associated RCK protein TrkA. *Nature* **496,** 317–322 (2013).
35. Roosild, T. P., Miller, S., Booth, I. R. & Choe, S. A mechanism of regulating transmembrane potassium flux through a ligand-mediated conformational switch. *Cell* **109,** 781–791 (2002).
36. Schlosser, A., Hamann, A., Bossemeyer, D., Schneider, E. & Bakker, E. P. NAD^+ binding to the *Escherichia coli* K^+-uptake protein TrkA and sequence similarity between TrkA and domains of a family of dehydrogenases suggest a role for NAD^+ in bacterial transport. *Mol. Microbiol.* **9,** 533–543 (1993).
37. Fisher, S. H. Regulation of nitrogen metabolism in *Bacillus subtilis*: vive la différence! *Mol. Microbiol.* **32,** 223–232 (1999).
38. Cortes, D. M., Cuello, L. G. & Perozo, E. Molecular architecture of full-length KcsA: role of cytoplasmic domains in ion permeation and activation gating. *J. Gen. Physiol.* **117,** 165–180 (2001).
39. Hodgkin, A. L. & Huxley, A. F. A quantitative description of membrane current and its applications to conduction and excitation in nerve. *J. Physiol. (Lond.)* **117,** 500–544 (1952).
40. Waters, C. M. & Bassler, B. L. Quorum sensing: cell-to-cell communication in bacteria. *Annu. Rev. Cell Dev. Biol.* **21,** 319–346 (2005).
41. Kato, S., Hashimoto, K. & Watanabe, K. Iron-oxide minerals affect extracellular electron-transfer paths of *Geobacter* spp. *Microbes Environ.* **28,** 141–148 (2013).
42. Pfeffer, C. *et al.* Filamentous bacteria transport electrons over centimetre distances. *Nature* **491,** 218–221 (2012).
43. Masi, E. *et al.* Electrical spiking in bacterial biofilms. *J. R. Soc. Interface* **12,** 1–10 (2014).
44. Pan, J. W. *et al.* Neurometabolism in human epilepsy. *Epilepsia* **49,** 31–41 (2008).
45. Petroff, O. A. C., Errante, L. D., Rothman, D. L., Kim, J. H. & Spencer, D. D. Glutamate-glutamine cycling in the epileptic human hippocampus. *Epilepsia* **43,** 703–710 (2002).
46. Meldrum, B. S. Glutamate as a neurotransmitter in the brain: review of physiology and pathology. *J. Nutr.* **130,** 1007S–1015S (2000).

Supplementary Information is available in the online version of the paper.

Acknowledgements We would like to thank S. Lockless, K. Süel, R. Wollman, T. Çağatay and M. Elowitz for comments during the writing of the manuscript, and C. Piggott for cloning help. A.P. is a Simons Foundation Fellow of the Helen Hay Whitney Foundation. J.G.-O. is supported by the Ministerio de Economia y Competitividad (Spain) and FEDER, under project FIS2012-37655-C02-01, and by the ICREA Academia Programme. This research was funded by the National Institutes of Health, National Institute of General Medical Sciences Grant R01 GM088428 and the National Science Foundation Grant MCB-1450867 50867 (both to G.M.S.). This work was also supported by the San Diego Center for Systems Biology (NIH Grant P50 GM085764).

Author Contributions G.M.S., A.P., J.L., M.A. and J.G.-O. designed the research, A.P. and J.L. performed the experiments, J.L. and A.P. performed the data analysis, J.G.O. performed the mathematical modelling, S.L. made the bacteria strains, and G.M.S., A.P., J.L. and J.G.-O. wrote the manuscript. All authors discussed the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to G.M.S. (gsuel@ucsd.edu).

METHODS

No statistical methods were used to predetermine sample size.

Strains. All experiments were done using *Bacillus subtilis* NCIB 3610. The wild-type strain was a gift from W. Winkler (University of Maryland)[47], and all other strains were derived from it and are listed in Extended Data Table 1. To make deletion strains, we used polymerase chain reaction (PCR) to amplify the desired regions from the wild-type strain. The PCR products were then put within the pER449 vector (gift from W. Winkler). For the *trkA* mutant, we deleted the C-terminal portion of *yugO* (amino acids 117–328), leaving only the N-terminal ion channel portion of *yugO* (amino acids 1–116). We identified the *trkA* region using Pfam (http://pfam.xfam.org/). All constructs were confirmed by direct sequencing and then integrated into the chromosome of the wild-type strain by a standard one-step transformation procedure[48]. Finally, chromosomal integrations were confirmed by colony PCR using the corresponding primers.

Growth conditions. The biofilms were grown in MSgg medium[16] which contains 5 mM potassium phosphate buffer (pH 7.0), 100 mM MOPS buffer (pH 7.0, adjusted using NaOH), 2 mM $MgCl_2$, 700 μM $CaCl_2$, 50 μM $MnCl_2$, 100 μM $FeCl_3$, 1 μM $ZnCl_2$, 2 μM thiamine HCl, 0.5% (v/v) glycerol and 0.5% (w/v) monosodium glutamate. The MSgg medium was made from stock solutions on the day of the experiment, and the stock solution for glutamate was newly made weekly.

Microfluidics. We followed methods similar to a previous study[18]. Briefly, we used the CellASIC ONIX Microfluidic Platform and the Y04D microfluidic plate (EMD Millipore). We used a pump pressure of 1 psi with only one media inlet open, which corresponds to a flow speed of $\sim$16 μm s^{-1}. On the day before the experiment, cells from −80 °C glycerol stock were streaked onto an LB agar plate and incubated at 37 °C overnight. The next morning, a single colony was picked from the plate and inoculated into 3 ml of LB broth and incubated in a 37 °C shaker. After 2.5 h of incubation, the cell culture was centrifuged at 2,100 relative centrifugal force (rcf) for 1 min, and the cell pellet was re-suspended in MSgg and immediately loaded into microfluidic chambers. After loading, cells in the microfluidic chamber were incubated at 37 °C for 90 min, and then the temperature was kept at 30 °C for the rest of the experiment.

Time-lapse microscopy. The growth of the biofilms was recorded using phase-contrast microscopy. The microscopes used were Olympus IX83 and DeltaVision PersonalDV. To image entire biofilms, 10 × objectives were used in most of the experiments. Biofilm phase contrast and fluorescence images were taken every 10 min, except in Fig. 2g where images were taken every 5 min. To generate Fig. 3b and Extended Data Figs 4a–c, where high temporal resolution was required, images were taken every minute. Whenever fluorescence images were recorded, we used the minimum exposure time that still provided a good signal-to-noise ratio (for example, we typically used 20 ms exposure for ThT and 100 ms exposure for APG-4).

Image analysis. Fiji/ImageJ (National Institutes of Health) and MATLAB (MathWorks) were used for image analysis. We generated custom scripts and used the image analysis toolbox to perform image segmentation on biofilm phase contrast images. To measure biofilm growth rate, we identified the biofilm area in each frame by segmenting the images and took the derivative of biofilm radius over time. We identified the radius by assuming circular growth of the colony and taking the length from the centre of the cell trap to the biofilm edge. To generate membrane potential curves, we measured the fluorescence of ThT within the biofilm using the ImageJ 'Plot Z-axis Profile' command and performed subsequent analysis, such as normalization and subtracting of baseline signal, in MATLAB.

Experimental reproducibility. Data shown in the main figures were drawn from a minimum of three independent experiments and often many more. For example, we analysed ThT oscillations (represented in Fig. 1c–e) in over 75 biofilms. In cases where only a single representative trace is shown, we analysed multiple regions within the biofilm to ensure accuracy of the analysis. In experiments comparing the wild-type and a mutant strain (*yugO* or *yugOΔtrkA*), we always performed head-to-head experiments (separate chambers in the same microfluidic device) on the same day using the same media to eliminate possible artefacts.

Mathematical modelling. The theoretical curves shown in Fig. 3d, e were generated using a mathematical model inspired by the Hodgkin–Huxley model of neuronal excitability[39] (Supplementary Information). The parameters used in the model (Extended Data Table 2) were constrained using a combination of literature values[24] and experimental data. Specifically, the response time to KCl shock (Fig. 3b) was used as a constraint on parameters with a time dimension and the spatial scale (lattice size of the 1d simulations) is extracted from the characteristic distance shown in Fig. 3g.

Theoretical estimate of potassium diffusion within biofilms. We used the diffusion coefficient of potassium in water (19.7×10^{-6} cm^2 s^{-1})[49] and reduced it to 70% in accordance with a reference on diffusion in biofilms[50], yielding the value of the diffusion coefficient (13.8×10^{-6} cm^2 s^{-1}) used in the mathematical model as well as the theoretical curves plotted alongside our experimental data. To estimate the rate of potassium propagation by diffusion, we used the formula for 2D mean squared displacement (MSD):

$$r = \sqrt{4Dt}$$

Where r is the displacement, D is the diffusion coefficient, and t is time. We used this relationship to generate the curve shown in Extended Data Fig. 3d. We directly compared the log–log slope of the experimental data (slope = 1.1, $R^2 = 0.96$) to that expected for diffusion (slope = 0.5) to further confirm that the experimental data cannot be explained by simple diffusion.

To estimate the decay of amplitude by diffusion, we used the formula for the concentration profile of 2D diffusion:

$$C(r,t) = \frac{M}{4\pi Dt}\exp\left(-\frac{r^2}{4Dt}\right)$$

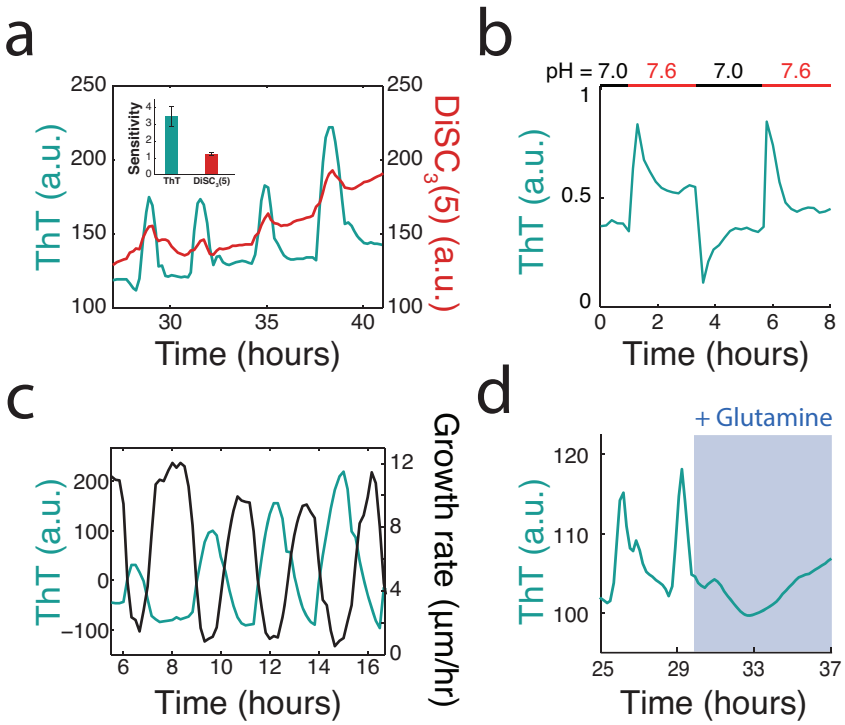
Where C is the concentration of potassium at displacement r and time t, M is a constant related to the initial pulse amplitude of potassium that we matched to the initial experimental pulse amplitude of APG-4, and D is the diffusion coefficient. We used this relationship to generate the curve in Fig. 2h.

Dyes and concentrations. Thioflavin T (ThT) and $DiSC_3(5)$ were used at 10 μM. We used ThT and $DiSC_3(5)$ to track relative changes in the membrane potential, where the fluorescence of ThT increases when the cell becomes more inside negative (hyperpolarizes). We found the sensitivity of ThT to be significantly higher than that of $DiSC_3(5)$, where sensitivity is defined as the ratio between the amplitude of oscillation and its error (Extended Data Fig. 1a). Furthermore, under our experimental conditions, $DiSC_3(5)$ appears to be absorbed by the PDMS in the microfluidic device. This hinders quantitative analysis (lower sensitivity) and also greatly increases the time required for the dye to diffuse into the biofilm.

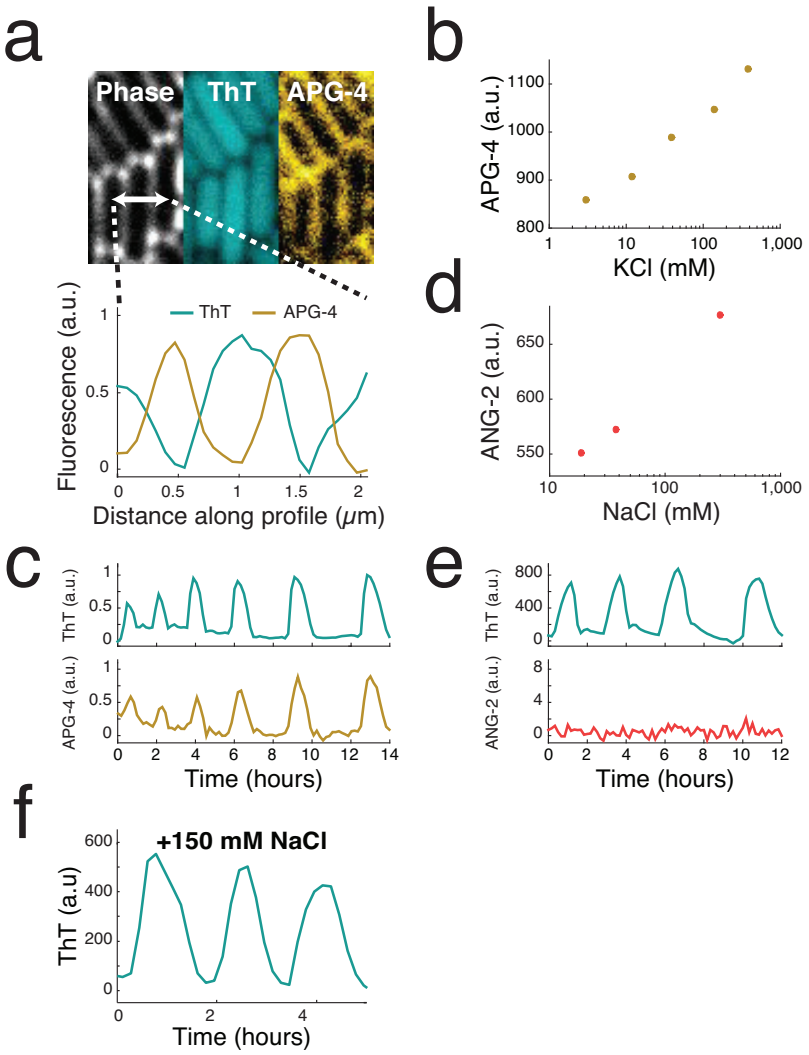
APG-4 (TEFLabs) was used at 2 μM. We used the membrane-impermeable TMA^+ salt form to track the extracellular concentration of potassium. We verified that APG-4 does not significantly diffuse into cells (Extended Data Fig. 2a). We also verified that APG-4 could measure extracellular potassium in MSgg media within our microfluidic device (Extended Data Fig. 2b).

We used ANG-2 (TEFLabs) at 2 μM. We used the membrane-impermeable TMA^+ salt form to track the extracellular concentration of sodium (Extended Data Fig. 3c, d).

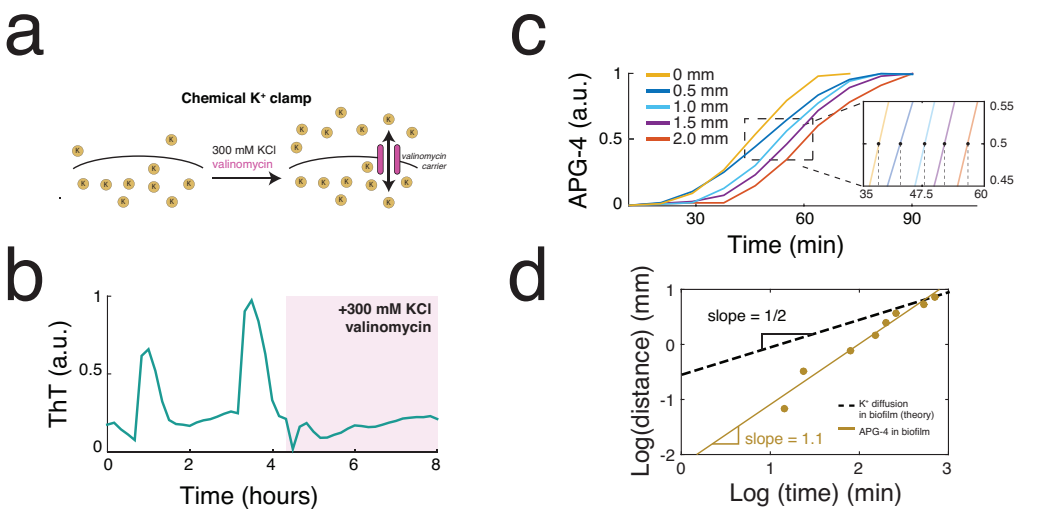
47. Irnov, I. & Winkler, W. C. A regulatory RNA required for antitermination of biofilm and capsular polysaccharide operons in Bacillales. *Mol. Microbiol.* **76,** 559–575 (2010).
48. Jarmer, H., Berka, R., Knudsen, S. & Saxild, H. H. Transcriptome analysis documents induced competence of Bacillus subtilis during nitrogen limiting conditions. *FEMS Microbiol. Lett.* **206,** 197–200 (2002).
49. Horvath, A. L. *Handbook of Aqueous Electrolyte Solutions: Physical Properties, Estimation and Correlation Methods* (Ellis Horwood Ltd, 1985).
50. Stewart, P. S. Diffusion in biofilms. *J. Bacteriol.* **185,** 1485–1491 (2003).



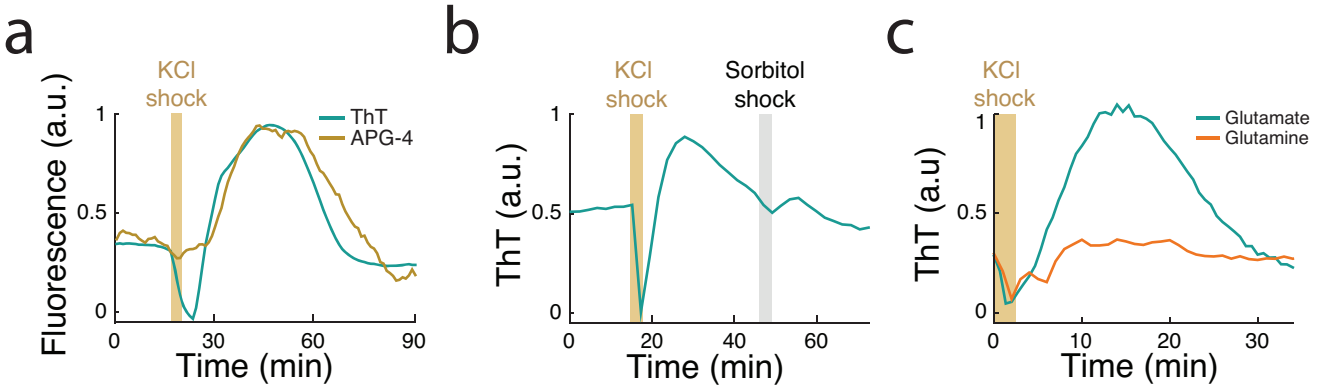
Extended Data Figure 1 | Thioflavin T (ThT) is a fluorescent reporter that is inversely related to the membrane potential. **a**, ThT and $DiSC_3(5)$, an established reporter of membrane potential in bacteria[23], both oscillate within biofilms. ThT has an approximately three fold higher sensitivity to changes in membrane potential compared to $DiSC_3(5)$. Sensitivity is defined as the ratio between peak height and error in peak height. Error bars indicate mean ± s.d. ($n = 8$ biofilm regions, averaged over the 4 pulses shown). **b**, The cellular ThT fluorescence depends on the external pH, where higher pH results in greater membrane potential, as expected[24]. ThT itself is insensitive to these pH changes and the traces are background subtracted to eliminate possible artefacts. Representative trace is selected from three independent biofilms. **c**, Oscillations in ThT and growth rate are inversely correlated, linking membrane potential oscillations to the metabolic cycle which produces periodic growth pauses[18]. Growth rate is calculated by taking the derivative of biofilm radius over time (Supplementary Information). Representative trace is selected from over 75 independent biofilms. **d**, Replacing glutamate with 0.2% glutamine, which eliminates the need to take up glutamate or retain ammonium, quenches ThT oscillations. This further suggests that ThT oscillations are specific to the metabolic cycle involving glutamate and ammonium. A representative trace was selected from three independent experiments.



Extended Data Figure 2 | A fluorescent reporter of extracellular potassium (APG-4) indicates that potassium has a role in membrane potential oscillations. **a**, High-resolution images showing the intracellular localization of ThT and primarily extracellular localization of APG-4 (top). Quantification of ThT and APG-4 along the 2 μm profile indicated in the phase image indicates that APG-4 does not significantly diffuse into the cell (bottom). Representative images are selected from six independent experiments. **b**, Induction curve for APG-4 generated using externally supplemented KCl. The experiment was repeated twice. **c**, Oscillations in extracellular potassium in the surrounding cell-free region during biofilm oscillations. These oscillations occurred during the experiment shown in Fig. 2b, c and the pulses are synchronized between the biofilm and the surrounding cell-free region. Representative trace is selected from six independent experiments. **d**, Induction curve for ANG-2 generated using externally supplemented NaCl. The experiment was repeated twice. **e**, Simultaneous measurement of ThT and ANG-2 indicates a lack of oscillations in extracellular sodium. Representative trace selected from three independent biofilms. **f**, Furthermore, perturbing extracellular sodium concentrations in the media had no detectable effect on membrane potential oscillations. A representative trace was selected from four independent experiments.



Extended Data Figure 3 | Active propagation of potassium signal within the biofilm. **a**, A chemical potassium clamp (300 mM KCl, matching the intracellular concentration[29], and 30 μM valinomycin) prevents the formation of potassium electrochemical gradients across the cellular membrane. Valinomycin is an antibiotic that creates potassium-specific carriers in the cellular membrane[32]. **b**, Clamping net potassium flux quenches oscillations in membrane potential. A representative trace was selected from two independent biofilms. **c**, Propagation of extracellular potassium is estimated by tracking the half-maximal position of the pulse over time. Representative traces are shown for a single pulse selected from one of six independent experiments. **d**, Propagation of extracellular potassium is relatively constant over time in contrast to diffusion that is expected to decay. The diffusion line is calculated using the mean squared displacement (MSD) and the diffusion coefficient for potassium in biofilms (Supplementary Information). Slopes are calculated from the same representative data shown in **c**.



Extended Data Figure 4 | External potassium affects the metabolic state of the cell. **a**, A potassium shock (300 mM KCl) produces an initial ThT decrease (depolarization) followed by a period of sustained ThT increase (hyperpolarization). ThT is inversely related to the membrane potential. A corresponding pulse in APG-4 during this ThT increase suggests that hyperpolarization is due to release of potassium. APG-4 signal due to the external potassium shock itself was subtracted using the cell-free background near the biofilm. A representative trace was selected from three independent experiments. **b**, ThT spikes in response to external potassium shock (300 mM KCl) but not an equivalent shock of 300 mM sorbitol, an uncharged solute. A representative trace was selected from three independent experiments. **c**, The hyperpolarization response occurs when cells are grown in glutamate but not when glutamate is replaced by 0.2% glutamine, which bypasses the need to take up glutamate or retain ammonium. A representative trace was selected from four independent biofilms.

Extended Data Table 1 | List of strains used in this study

Strain	Genotype	Source
Wild type	*B. subtilis* NCIB 3610	*1*
ΔyugO	*yugO::* neo	This study
yugOΔtrkA	*trkA::* neo	This study

Extended Data Table 2 | Parameter values used in the model

Parameter	Value
g_K	30 min^{-1}
g_L	0.2 min^{-1}
V_{K0}	-380 mV
V_{L0}	-156 mV
S_{th}	40 μM
V_{th}	-150 mV
α_0	2 min^{-1}
β	1.3 min^{-1}
m	1
F	5.6 mM/mV

Parameter	Value
σ	0.2 mV
δ_K	1 mV/mM
δ_L	8 mV/mM
γ_s	0.1 min^{-1}
γ_e	10 min^{-1}
γ_t	4 min^{-1}
α_s	1 μM/(min mV)
α_t	1 μM/(min mV)
D	$13.8\times10^{-6}cm^2/s$

ARTICLE

doi:10.1038/nature15247

Architecture of the mammalian mechanosensitive Piezo1 channel

Jingpeng Ge[1,2]*, Wanqiu Li[2]*, Qiancheng Zhao[1,3]*, Ningning Li[2]*, Maofei Chen[1,2], Peng Zhi[3], Ruochong Li[1,2], Ning Gao[2], Bailong Xiao[1,3,4] & Maojun Yang[1,2]

Piezo proteins are evolutionarily conserved and functionally diverse mechanosensitive cation channels. However, the overall structural architecture and gating mechanisms of Piezo channels have remained unknown. Here we determine the cryo-electron microscopy structure of the full-length (2,547 amino acids) mouse Piezo1 (Piezo1) at a resolution of 4.8 Å. Piezo1 forms a trimeric propeller-like structure (about 900 kilodalton), with the extracellular domains resembling three distal blades and a central cap. The transmembrane region has 14 apparently resolved segments per subunit. These segments form three peripheral wings and a central pore module that encloses a potential ion-conducting pore. The rather flexible extracellular blade domains are connected to the central intracellular domain by three long beam-like structures. This trimeric architecture suggests that Piezo1 may use its peripheral regions as force sensors to gate the central ion-conducting pore.

Mechanosensitive cation channels have key roles in converting mechanical stimuli into various biological activities, such as touch, hearing and blood pressure regulation, through a process termed mechanotransduction[1]. Piezo proteins have recently been identified as pore-forming subunits of the long-sought-after mechanosensitive cation channels in metazoans[2–8]. A single fly *Piezo* gene has been shown to be involved in mechanical nociception[8]. There are two Piezo proteins in vertebrates: Piezo1 and Piezo2. In vertebrates, including fish[9], birds[10], rodents[11–14] and humans[15], Piezo2 mediates gentle touch sensation. By contrast, Piezo1 has broad roles in multiple physiological processes, including sensing shear stress of blood flow for proper blood vessel development[16,17], regulating red blood cell function[18,19] and controlling cell migration and differentiation[20,21]. In humans, mutations of *PIEZO1* or *PIEZO2* have been linked to several genetic diseases, including dehydrated hereditary stomatocytosis[22–27], distal arthrogryposis type 5 (ref. 28), Gordon syndrome and Marden–Walker syndrome[29]. These findings demonstrate the functional importance of Piezo channels, as well as their pathological relevance and potential as therapeutic targets.

Despite the functional importance of Piezo proteins, their gating mechanisms and three-dimensional (3D) structures are yet to be defined. They do not bear notable sequence and structural homology to any known classes of ion channel, such as voltage- or ligand-gated channels[30–32], transient receptor potential (TRP) channels[33,34], prokaryotic mechanosensitive channels[35–38] or eukaryotic mechanosensitive two-pore-domain potassium channels[39]. Mammalian Piezo proteins contain more than 2,500 residues with numerous predicted transmembrane segments[2,3,7,40] and form homo-oligomerized channel complexes[3]. However, the exact stoichiometry, topology, architecture and functional domains involved in pore formation, force sensing and regulation remain to be solved.

Combining protein engineering, X-ray crystallography, single-particle cryo-electron microscopy and live-cell immunostaining, we have obtained the medium-resolution structure of the full-length Piezo1 channel. Our results provide key insights into the ion-conducting and gating mechanisms of this novel class of mechanosensitive ion channels.

Piezo1 forms a homotrimer

Our initial effort was focused on obtaining a sufficient amount of acceptably homogenous Piezo proteins. Human, mouse and *Drosophila* Piezo complementary DNAs, in full-length or truncated forms, were cloned into a vector encoding a carboxy-terminal (C-terminal) glutathione *S*-transferase (GST) tag with a precision protease cleavage site in between (Piezo1–pp–GST). Constructs were tested for their expression using transient transfection in HEK293T cells. A large number of detergents in various classes were screened for their compatibility with the extraction and purification of Piezo proteins. Finally, a combination of mouse Piezo1 with the detergent C12E10 was used for purification and structural determination.

Gel filtration chromatography showed that Piezo1–pp–GST and Piezo1 without the GST tag both contained two forms of oligomer, but at different ratios (Fig. 1a–c and Extended Data Fig. 1). On native gels, Piezo1–pp–GST migrated as a major band at a molecular weight of about 1,200 kDa and a minor one at about 900 kDa (Fig. 1c). This result seemed consistent with a previous study, which suggested that Piezo1 fused to GST formed a homotetramer[3]. However, examination of Piezo1–pp–GST proteins by negative-staining electron microscopy showed an ostensibly dimeric arrangement of particles (Fig. 1d, e). Two-dimensional (2D) classification of these particles indicated that the two halves were highly similar (Fig. 1f), suggesting that the dimerized GST tag may mediate further dimerization of Piezo1 complexes. Consistent with this possibility, Piezo1 with the GST tag cleaved displayed mainly a molecular weight of 900 kDa on native gels (Fig. 1c). Moreover, almost no particles with the dimeric arrangement could be observed in the tag-free Piezo1 sample. Rather, particles with a three-fold symmetry were clearly detected (Fig. 1g–i). As a further confirmation, Flag-tagged Piezo1 displayed a major band at about 900 kDa on native gels (Fig. 1c). Thus, our data suggest that the major oligomeric state of the purified Piezo1 is trimeric. The majority of

[1]Tsinghua-Peking Joint Center for Life Sciences, School of Life Sciences or Medicine, Tsinghua University, Beijing 100084, China. [2]Ministry of Education, Key Laboratory of Protein Sciences, School of Life Sciences, Tsinghua University, Beijing 100084, China. [3]Department of Pharmacology and Pharmaceutical Sciences, School of Medicine, Tsinghua University, Beijing 100084, China. [4]IDG/McGovern Institute for Brain Research, Tsinghua University, Beijing 100084, China.
*These authors contributed equally to this work.

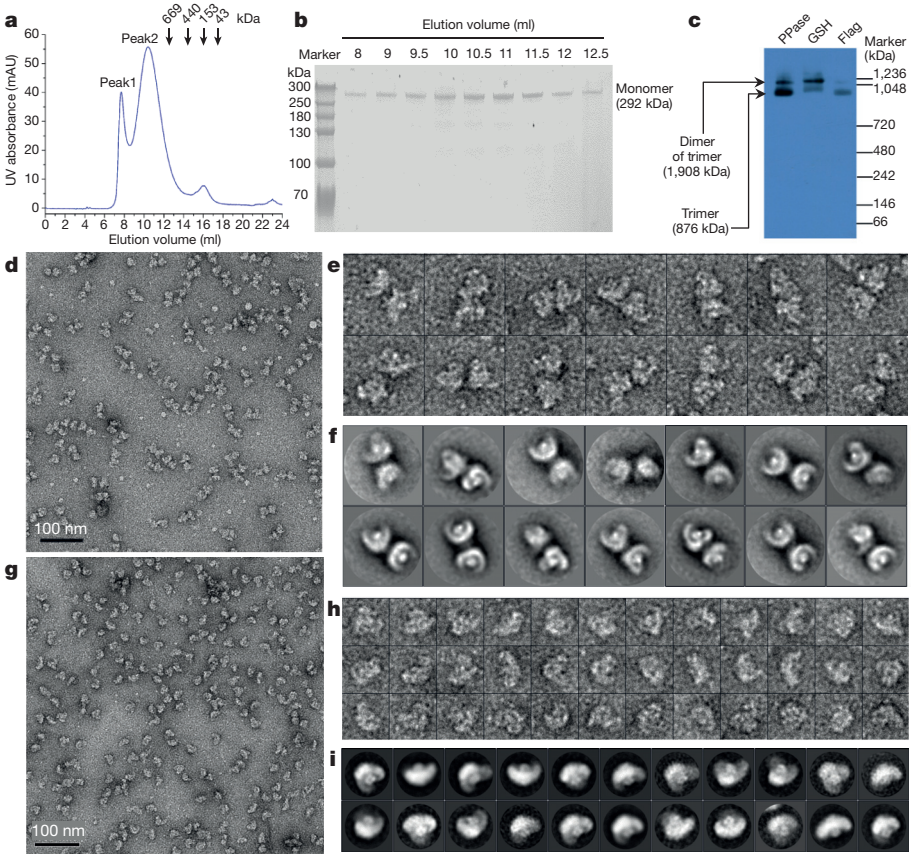


Figure 1 | Piezo1 forms a homotrimer. **a**, A representative trace of gel filtration of the full-length Piezo1, with molecular weight markers indicated. UV, ultraviolet. **b**, Protein samples of the indicated fractions were subjected to SDS–PAGE and Coomassie blue staining. **c**, Native gel and western blotting analysis of GST-cleaved Piezo1 (PPase), Piezo1–pp–GST (GSH) and Piezo1–Flag (Flag) samples with an anti-Piezo1 antibody. **d**, A representative micrograph of the negatively stained Piezo1–pp–GST. **e**, Raw particles of Piezo1–pp–GST. **f**, 2D class averages of Piezo1–pp–GST particles. **g**, A representative micrograph of the negatively stained Piezo1. **h**, Raw particles of Piezo1. **i**, 2D class averages of Piezo1 particles.

Piezo1–pp–GST fusion proteins form a dimer of trimers, as a result of the dimerized GST tags.

The unusual migration of the 1,900-kDa Piezo1–pp–GST dimer of trimers near the 1,200-kDa marker might have led to the incorrect characterization of Piezo1–pp–GST as a tetramer in the previous report[3]. The large native size of the protein, together with its numerous transmembrane segments, might have resulted in its unusual mobility on native gels owing to the influence of the detergents. Nevertheless, we could not completely exclude the possibility that Piezo1 exists in other oligomeric states on the membrane or under different conditions *in vitro*, a scenario observed in previous studies of other ion channels (for example, Orai channels)[41,42].

Three-blade, propeller-shaped Piezo1 homotrimer

Using a single-particle approach during cryo-electron microscopy, we determined the trimeric structure of Piezo1 (Fig. 2a–d and Extended Data Figs 2–5). Notably, the density map revealed that Piezo1 formed a three-blade, propeller-shaped architecture, with distinct regions resembling the typical structural components of a propeller, including three blades and a central cap. Viewed from the top, the diameter and the axial height of the structure are 200 Å and 155 Å, respectively (Fig. 2d). The transmembrane region could be readily located and contains many paired density rods, in good agreement with the 2D analyses (Fig. 2c–f). The transmembrane region contains three extended and twisted arrays of transmembrane helices (Fig. 2f, second from left). Beyond the transmembrane helical array, three thick distal blades are arranged in a superhelical fashion and each blade also has a helicoidal surface (Fig. 2d, e and f, second from right). A single central cap sits above the surface of the transmembrane core with a gap (~8 Å) in between (Fig. 2e). Furthermore, a tightly packed region, likely to be a compact soluble domain, is located on the opposite side of the cap, right below the transmembrane region (Fig. 2e). Three long, distinct density rods exposed on the outer surface of the transmembrane region, hereafter termed beam, seem to connect the distal end of the transmembrane region and the blades mechanically to the centre of the trimeric complex at the bottom face. The diameter of the density rod suggests that the beam is composed of a two-stranded coiled coil (Fig. 2d, e).

Topology determination

The proposed detachment of the cap from the transmembrane core indicates that it is likely to be a soluble region. A topological prediction model suggests that residues from 2210 to 2457 (termed the C-terminal extracellular domain, CED) constitute a large extracellular loop followed by the last transmembrane segment at the C terminus[43]. To test whether this region constitutes the cap, we constructed and purified the deletion-mutant Piezo1(Δ2219–2453) and examined it by negative-staining electron microscopy. 2D classification of Piezo1(Δ2219–2453) particles revealed the central cap was absent in 2D class averages (Extended Data Fig. 6a, b), confirming that this region indeed forms the cap.

Next, we solved the crystal structure of the CED (Piezo1(2214–2457)) (Fig. 2g and Extended Data Table 1), which was similar to that of the same region of *Caenorhabditis elegans* Piezo reported recently[43]. The root-mean-square deviation of 181 aligned α-carbon atoms between these two structures is 1.7 Å (Extended Data Fig. 6c, d). The amino (N) and C termini of the CED are on the same side and close to each other (Fig. 2g), consistent with the topological prediction[40,43] that the CED is located between the last two transmembrane segments in the C-terminal region of Piezo1.

The CED formed a trimer in both gel filtration and crystal lattice (Extended Data Fig. 6d, e). A direct and rigid fitting of the crystallographic trimer of the CED into the cryo-electron microscopy density map resulted in a match, with a correlation coefficient of 0.89 (Extended Data Fig. 6f). These results demonstrate that the cap is formed by a CED trimer, further supporting the conclusion that the full-length Piezo1 forms a homotrimer. Furthermore, the high consistency between the crystal structure and the cryo-electron

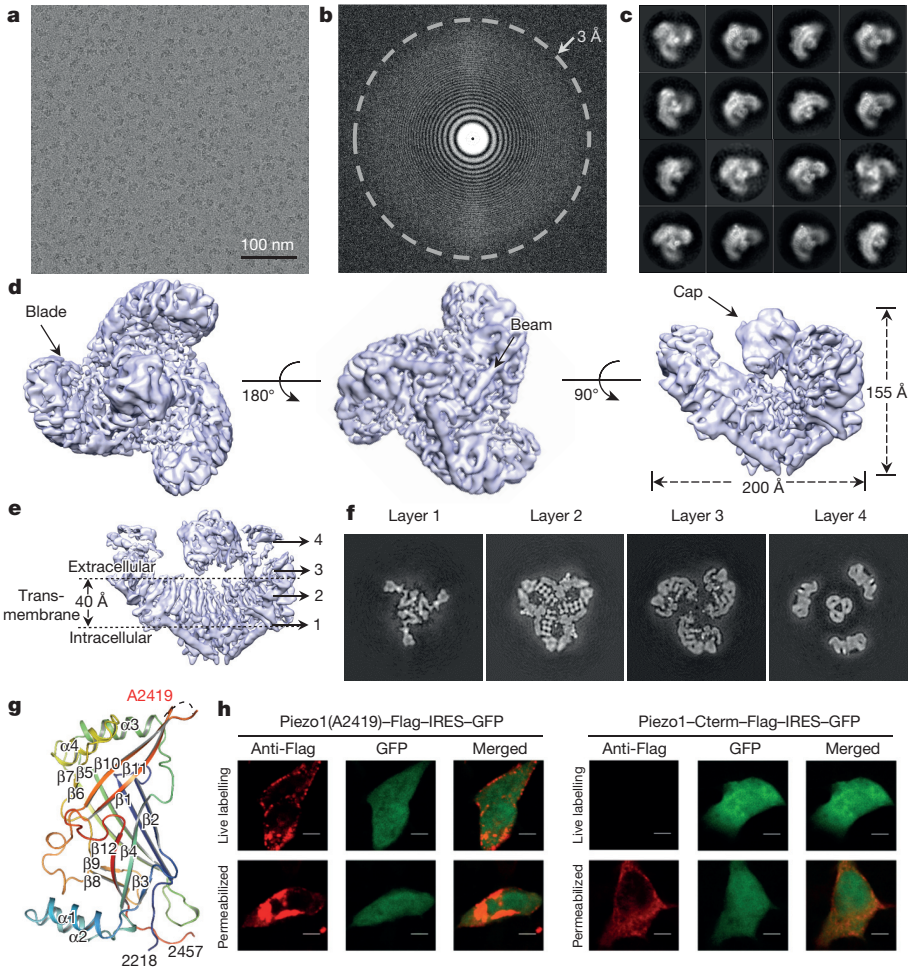


Figure 2 | Overall structure of Piezo1. **a**, A representative cryo-electron micrograph of Piezo1. **b**, Power spectrum of the micrograph in **a**, with the 3-Å frequency indicated. **c**, Representative 2D class averages of Piezo1 particles, showing fine features of the trimeric complex. **d**, Top, bottom and side views of an unsharpened map (5σ contour level) of Piezo1, with distinct regions labelled. The dimensions of the trimeric structure is shown in the rightmost panel. **e**, Side view of the sharpened map (6σ contour level) of Piezo1 filtered to a resolution of 4.8 Å, with the transmembrane region indicated. **f**, Selected *z*-slices of the final sharpened map corresponding to the layers indicated by the numbered arrows in **e**. **g**, The cartoon model of the crystal structure of a single C-terminal extracellular domain. The dashed line indicates the missing residues. The Flag tag was inserted after residue A2419. **h**, Immunostaining of cells transfected with the indicated constructs with an anti-Flag antibody either in live labelling (top row) or after fixation and permeabilization (bottom row). Scale bars, 10 μm. GFP, green fluorescent protein; IRES, internal ribozyme entry site.

microscopy map of the cap domain confirmed the correctness of the density map and determined the handedness of the map.

To further confirm the topological location of the CED and the C terminus of Piezo1, we performed immunolabelling of live HEK293T cells expressing Piezo1 with a Flag tag fused either in a flexible loop of the CED (after A2419) or at the C terminus of Piezo1. Using confocal microscopy, we found that the Flag tag could be labelled on the plasma membrane of live cells only when inserted in the CED and not at the C terminus (Fig. 2h). These data demonstrate that the CED is an extracellular domain, whereas the C terminus is intracellular, consistent with a recent report[40]. Consequently, this suggests that both the central cap and the three blades locate at the extracellular side, whereas the beams locate at the intracellular side.

The transmembrane skeleton

Piezo proteins have been predicted to contain an unusually large number of transmembrane segments (about 30–40) in one molecule[2,3,7,40]. Several potential topology models of Piezo have recently been proposed, with the number of transmembrane segments ranging from 10 to 38 (ref. 40). The local resolution of the cryo-electron microscopy density map shows that the transmembrane region is associated with a higher resolution, which allowed us to build a *de novo* alanine model with 492 amino acids for the more readily identified transmembrane segments, beam and the intracellular C-terminal domain (CTD). Together with the 227 amino acids of the CED, we built a total of 719 residues (out of 2,547 amino acids) for each monomer (Fig. 3a and Extended Data Figs 7, 8). The whole transmembrane skeleton displays a three-winged arrangement, with each extended wing being slightly twisted (Fig. 3b). From the map, 14 transmembrane segments could be readily recognized on each wing. A potential topology of at least 14 transmembrane segments for each protomer is consistent with a recent topology model of 18 transmembrane segments, instead of 38 transmembrane segments[40]. In line with this observation, a single blade has a volume comparable to the cap region, which is made up of about 700 residues. Thus, some of the predicted N-terminal helices should reside in the distal extracellular regions.

To facilitate the description of our structure and based on known features of ion channels, we refer to the core transmembrane segments as inner helix (IH) and outer helix (OH) and to the peripheral transmembrane arrays as peripheral helix (PH) (Fig. 3). The 12 PHs from the same monomer are organized as six helical pairs, extending from the central axis to the periphery of the complex (Fig. 3b). They are connected to the extracellular blade. The density for the connecting sequences from PH1 to PH7 allowed us to make tentative connections between them, except for the connection between PH4 and PH5 (Extended Data Fig. 8a).

Main-chain tracing of the PH1, IH and OH towards the transmembrane core in the density map, together with the information from topology (Fig. 3c) and secondary structure prediction (Extended Data Fig. 9), allowed us to map these three transmembrane segments on the primary sequence and assign some of the linker sequences between them into the corresponding density features. These analyses suggest that the OH connects to PH1 through four continuous α-helices, which form a unique hairpin structure at the interface of two adjacent subunits. This hairpin structure, termed the anchor, penetrates into the inner leaflet of the membrane, with a long helix ($\alpha4^{anchor}$) roughly parallel to the membrane (Fig. 3a, right and Extended Data Fig. 8b). The remaining density features in the map include the IH and its connecting density (also four α-helices) all the way to the intracellular surface of the channel, suggesting that the IH is the last

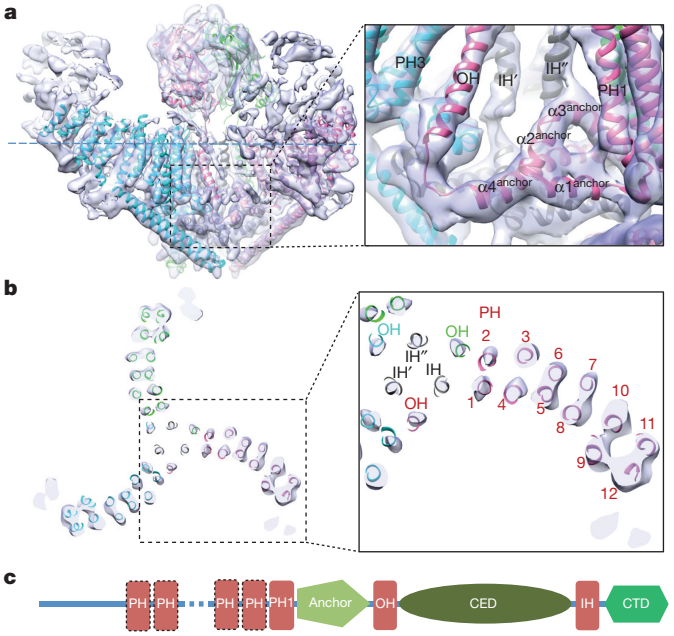


Figure 3 | Organization of the transmembrane skeleton. **a**, A side view of the cryo-electron microscopy density map superimposed with separately coloured poly-alanine models of each subunit. The boxed region is enlarged to illustrate the anchor domain. **b**, A *z*-slice representation of the overall organization of the transmembrane skeleton of the layer indicated by the blue dashed line in **a**. The boxed region is amplified to illustrate the central transmembrane core that consists of three IHs and three OHs and wings of the peripheral helices (PH1–PH12). Owing to the ambiguity in the connection, the three IHs are not assigned to each subunit and thus labelled as IH, IH′ and IH″. **c**, The model represents the topology of the C-terminal part of Piezo1. Different structural units are indicated.

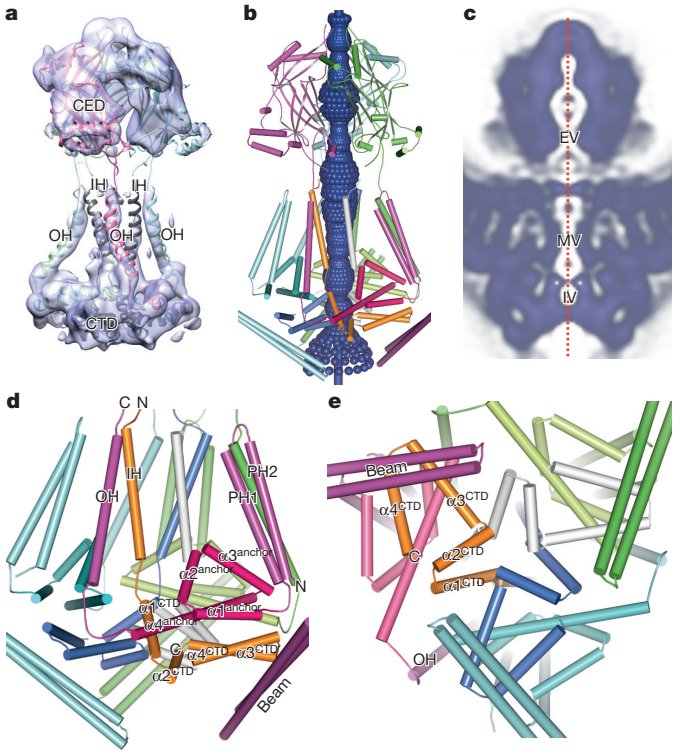


Figure 4 | Putative ion-conducting pore. **a**, Surface representation (transparent) of the segmented map of the putative pore module, including the OH, CED, IH and CTD. **b**, Same as a, but the model is superimposed with the putative ion-conducting pore (deep blue), produced by HOLE[48] with the poly-alanine model and the CED crystal structure. **c**, Central slice of the rotationally averaged density map, highlighting a continuous central pore along the *z*-axis (red dotted line). The extracellular vestibule (EV), transmembrane vestibule (MV) and intracellular vestibule (IV) regions are labelled. **d**, A side view of the CTD and the pore module consisting of the OH, IH and the CTD helices. **e**, Same as d, but viewed from the intracellular side.

transmembrane segment from the C terminus. In line with this assignment, the intracellular C terminus is located at the centre of the intracellular side, as indicated by the location of the C-terminal GST tag in Piezo1–pp–GST.

Together with the finding that the CED is inserted between the last two transmembrane segments from the C terminus, the OH is likely to be the second-to-last transmembrane segment from the C terminus, because of the close distance (matching the length of the linker sequences) between the N terminus of the CED and the extracellular end of the OH (Fig. 3a and Extended Data Fig. 9). In addition, the distance constraint enabled us to put a connection between a specific OH and one of the three N termini of the CED domain. However, we cannot unambiguously connect a specific IH to the three possible C termini of the CED.

Nevertheless, with the primary sequence of the PH1–anchor–OH–CED from one monomer fixed in the density map, a clear separation of the three subunits on the 3D structure could be achieved (Fig. 3a). The presence of the anchor domain also seems to result in a clockwise swapping of the OH–CED of one monomer (viewed from the cap) into a region of the neighbouring monomer. This helix-swapping arrangement might be critical for the stabilization of the Piezo1 trimer. Although unambiguous sequence assignment at the residue level was not feasible, this anchor domain of Piezo1 could be mapped to residues around 2100 to 2190, a region containing the most evolutionarily conserved sequence motif, PF(X2)E(X6)W (2129–2140), among Piezo homologues (Extended Data Fig. 9)[44]. The disease-causing mutation Piezo1(T2142) (T2127 of PIEZO1 in humans)[23] is located in this region, supporting the functional relevance of the anchor. Another mutation targeting this motif, Piezo1(E2133), was found to affect the Piezo1 channel pore properties[40].

Each wing of the transmembrane region sits on a coiled-coil beam exposed at the intracellular surface. The beam is about 80 Å in length and positioned at about 30° relative to the membrane (Fig. 3a). It originates peripherally at the intracellular side of the PH7–PH8 pair and ends near the central axis of the trimer, where it seems to interact with the anchor and the CTD (Fig. 3a). This organization suggests that the three beams might be responsible for transmitting conformational changes from peripheral transmembrane segments and the extracellular blades to the central region, where the ion-conducting pore is most likely to reside.

The ion-conducting pore

The centre of the Piezo1 channel within the membrane consists of six transmembrane helices in a triangular arrangement (Fig. 3b, right and Fig. 4). Three IHs, presumably extended from the C termini of the CEDs, are located at the innermost position and seem to line a central pore. Three OHs, extended from the N termini of the CEDs, further enclose the three IHs (Fig. 4a). This central region, including the IH–OH pairs, the CEDs and the CTDs, probably comprises the pore module of Piezo1. The lack of side-chain information in the three IHs prevented us from accurately determining the radius of the pore. Nonetheless, apparent restriction sites could be readily detected, suggesting that they are potential gating positions. The central slice of the rotationally averaged density map revealed a continuous central channel along the *z*-axis, including an extracellular vestibule within the cap, a transmembrane vestibule enclosed by the three IHs and an intracellular vestibule formed by the trimeric CTD (Fig. 4b–e). The organization of the central transmembrane core and the pore is reminiscent of the trimeric $P2X_4$ channels[32] and acid-sensing ion channels[45], although they possess only two transmembrane helices

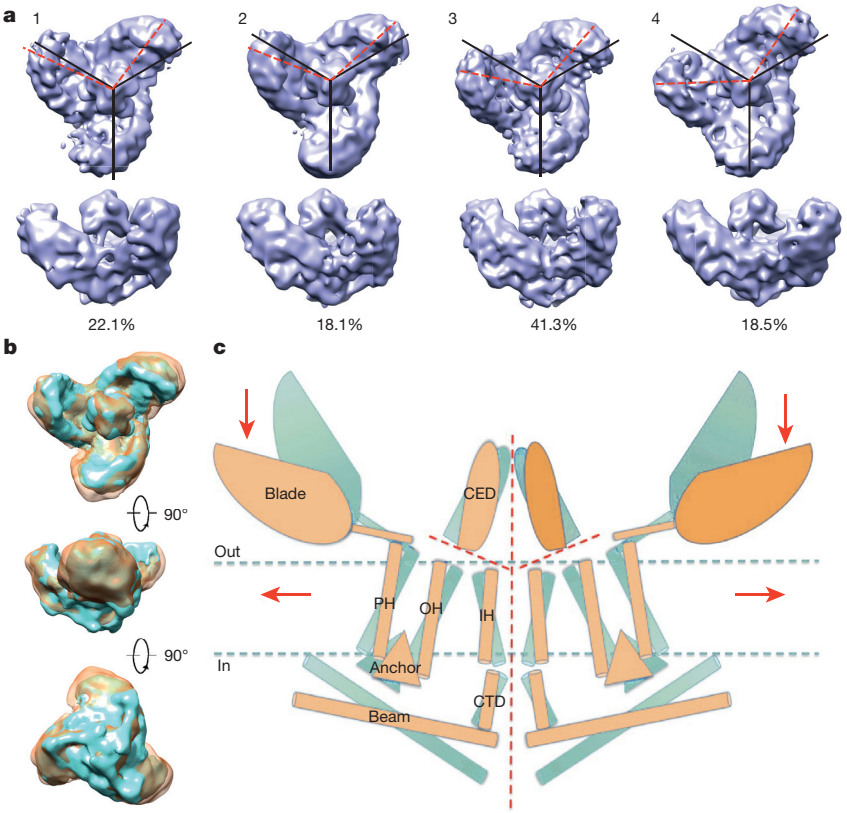


Figure 5 | Conformational heterogeneity of the 'blade' and a proposed model of force-induced gating of Piezo channels. **a**, Representative classes of Piezo1 structures from symmetry-free 3D classification. For each top-viewed structure, three black lines (120° interval) are drawn to illustrate the expected position of blades on the basis of perfect C3 symmetry. Red dashed lines represent observed positions of the blades. **b**, Structural comparison between further-refined maps of structures 4 (orange) and 3 (cyan) in **a**, showing the centripetal movement of the blades (top) and the tilted movement of the beams relative to the plasma membrane plane (bottom). **c**, Proposed model of the force-induced gating of Piezo channels. The blue and orange models represent the closed and open state channels, respectively. Red dashed lines indicate the possible ion-conduction pathways. Presumably, force-induced motion (red arrows) of the peripheral blade or PHs leads to conformational arrangement and gating of the channel.

and a large extracellular domain in each monomer. Based on this structural information, we propose that the OH–CED–IH–CTD-containing region functions as the pore module of Piezo channels (Fig. 4). According to our assignment, this pore module comprises the C-terminal region from residues 2172 to 2547. This is consistent with a recent study showing that the portion from 1974 to the C terminus of Piezo1 is essential for ion permeation properties[40].

The flexible blades as potential force sensors

The local resolution map shows that the three blades of Piezo1 have smeared densities at their distal ends and fragmented density in the sharpened map (Fig. 2d, e). In contrast, the cap, transmembrane skeleton, beam and CTD are better defined and display apparent secondary structural features. The blades of Piezo1 are highly flexible (Figs 2c, 3a and Extended Data Fig. 8). Indeed, comparison of different classes of the structures from symmetry-free 3D classification reveals several motion modes for the blade (Fig. 5a, b and Extended Data Fig. 5a). The most notable one is that the rotational spacing between two adjacent blades varies from 100° to 140° (Fig. 5a). Other less pronounced but identifiable conformational variations include the tilting of the blade relative to the plasma membrane and curvature changes on the helicoidal surface (Fig. 5b). Further supporting the structural flexibility of the blade regions, subregion refinement (see Methods) considerably improved the densities of the cap, but not that of the blade. The large conformational heterogeneity in the blades could be the main factor hampering high-resolution structural refinement of the entire structure. However, the structural flexibility of the propeller-like blades could be functionally meaningful. For example, they might serve as sensors of mechanical force exerted on the channel, thus contributing to mechanical gating of Piezo1 (Fig. 5c).

The recently resolved cryo-electron microscopy structure of human TRPA1 reveals a fourfold propeller-like structure composed of numerous ankyrin repeats[33]. Although TRPA1 alone is not sufficient to mediate mechanosensitive currents, it has been proposed to mediate slowly adapting mechanically activated currents in somatosensory neurons[46,47], raising an intriguing possibility that TRPA1 may employ the propeller-like structure to confer mechanosensitivity under certain circumstances. It remains possible that other extracellular or intracellular proteins may interact with and regulate Piezo channels. These hypotheses merit further investigation.

Conclusions

The medium-resolution cryo-electron microscopy structure of Piezo1 provides critical insights into the general architecture, oligomerization state and topological organization of Piezo channels. Our putative assignment of the central ion-conducting pore, mechanosensing and transduction components serves as a testable framework for dissection of the structure and mechanism of this class of channels.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 26 March; accepted 24 July 2015.
Published online 21 September, corrected online 4 November 2015 (see full-text HTML version for details).

1. Chalfie, M. Neurosensory mechanotransduction. *Nature Rev. Mol. Cell Biol.* **10,** 44–52 (2009).
2. Coste, B. *et al.* Piezo1 and Piezo2 are essential components of distinct mechanically activated cation channels. *Science* **330,** 55–60 (2010).
3. Coste, B. *et al.* Piezo proteins are pore-forming subunits of mechanically activated channels. *Nature* **483,** 176–181 (2012).
4. Nilius, B. & Honoré, E. Sensing pressure with ion channels. *Trends Neurosci.* **35,** 477–486 (2012).
5. Volkers, L., Mechioukhi, Y. & Coste, B. Piezo channels: from structure to function. *Pflugers Arch.* **467,** 95–99 (2015).
6. Bae, C., Gottlieb, P. A. & Sachs, F. Human PIEZO1: removing inactivation. *Biophys. J.* **105,** 880–886 (2013).
7. Gottlieb, P. A. & Sachs, F. Piezo1: properties of a cation selective mechanical channel. *Channels* **6,** 214–219 (2012).
8. Kim, S. E., Coste, B., Chadha, A., Cook, B. & Patapoutian, A. The role of *Drosophila* Piezo in mechanical nociception. *Nature* **483,** 209–212 (2012).
9. Faucherre, A., Nargeot, J., Mangoni, M. E. & Jopling, C. *piezo2b* regulates vertebrate light touch response. *J. Neurosci.* **33,** 17089–17094 (2013).
10. Schneider, E. R. *et al.* Neuronal mechanism for acute mechanosensitivity in tactile-foraging waterfowl. *Proc. Natl Acad. Sci. USA* **111,** 14941–14946 (2014).

11. Maksimovic, S. *et al.* Epidermal Merkel cells are mechanosensory cells that tune mammalian touch receptors. *Nature* **509,** 617–621 (2014).
12. Woo, S. H. *et al.* Piezo2 is required for Merkel-cell mechanotransduction. *Nature* **509,** 622–626 (2014).
13. Ranade, S. S. *et al.* Piezo2 is the major transducer of mechanical forces for touch sensation in mice. *Nature* **516,** 121–125 (2014).
14. Ikeda, R. *et al.* Merkel cells transduce and encode tactile stimuli to drive Aβ-afferent impulses. *Cell* **157,** 664–675 (2014).
15. Schrenk-Siemens, K. *et al.* PIEZO2 is required for mechanotransduction in human stem cell-derived touch receptors. *Nature Neurosci.* **18,** 10–16 (2015).
16. Ranade, S. S. *et al.* Piezo1, a mechanically activated ion channel, is required for vascular development in mice. *Proc. Natl Acad. Sci. USA* **111,** 10347–10352 (2014).
17. Li, J. *et al.* Piezo1 integration of vascular architecture with physiological force. *Nature* **515,** 279–282 (2014).
18. Faucherre, A., Kissa, K., Nargeot, J., Mangoni, M. E. & Jopling, C. Piezo1 plays a role in erythrocyte volume homeostasis. *Haematologica* **99,** 70–75 (2014).
19. Cahalan, S. M. *et al.* Piezo1 links mechanical forces to red blood cell volume. *eLife* **4,** 07370 (2015).
20. McHugh, B. J. *et al.* Integrin activation by Fam38A uses a novel mechanism of R-Ras targeting to the endoplasmic reticulum. *J. Cell Sci.* **123,** 51–61 (2010).
21. Pathak, M. M. *et al.* Stretch-activated ion channel Piezo1 directs lineage choice in human neural stem cells. *Proc. Natl Acad. Sci. USA* **111,** 16148–16153 (2014).
22. Albuisson, J. *et al.* Dehydrated hereditary stomatocytosis linked to gain-of-function mutations in mechanically activated PIEZO1 ion channels. *Nature Commun.* **4,** 1884 (2013).
23. Andolfo, I. *et al.* Multiple clinical forms of dehydrated hereditary stomatocytosis arise from mutations in *PIEZO1*. *Blood* **121,** 3925–3935 (2013).
24. Bae, C., Gnanasambandam, R., Nicolai, C., Sachs, F. & Gottlieb, P. A. Xerocytosis is caused by mutations that alter the kinetics of the mechanosensitive channel PIEZO1. *Proc. Natl Acad. Sci. USA* **110,** E1162–E1168 (2013).
25. Beneteau, C. *et al.* Recurrent mutation in the PIEZO1 gene in two families of hereditary xerocytosis with fetal hydrops. *Clin. Genet.* **85,** 293–295 (2014).
26. Shmukler, B. E. *et al.* Dehydrated stomatocytic anemia due to the heterozygous mutation R2456H in the mechanosensitive cation channel PIEZO1: a case report. *Blood Cells Mol. Dis.* **52,** 53–54 (2014).
27. Zarychanski, R. *et al.* Mutations in the mechanotransduction protein PIEZO1 are associated with hereditary xerocytosis. *Blood* **120,** 1908–1915 (2012).
28. Coste, B. *et al.* Gain-of-function mutations in the mechanically activated ion channel PIEZO2 cause a subtype of distal arthrogryposis. *Proc. Natl Acad. Sci. USA* **110,** 4667–4672 (2013).
29. McMillin, M. J. *et al.* Mutations in *PIEZO2* cause Gordon syndrome, Marden–Walker syndrome, and distal arthrogryposis type 5. *Am. J. Hum. Genet.* **94,** 734–744 (2014).
30. Zhang, X. *et al.* Crystal structure of an orthologue of the NaChBac voltage-gated sodium channel. *Nature* **486,** 130–134 (2012).
31. Payandeh, J., Scheuer, T., Zheng, N. & Catterall, W. A. The crystal structure of a voltage-gated sodium channel. *Nature* **475,** 353–358 (2011).
32. Kawate, T., Michel, J. C., Birdsong, W. T. & Gouaux, E. Crystal structure of the ATP-gated P2X(4) ion channel in the closed state. *Nature* **460,** 592–598 (2009).
33. Paulsen, C. E., Armache, J. P., Gao, Y., Cheng, Y. & Julius, D. Structure of the TRPA1 ion channel suggests regulatory mechanisms. *Nature* **520,** 511–517 (2015).
34. Liao, M., Cao, E., Julius, D. & Cheng, Y. Structure of the TRPV1 ion channel determined by electron cryo-microscopy. *Nature* **504,** 107–112 (2013).
35. Liu, Z., Gandhi, C. S. & Rees, D. C. Structure of a tetrameric MscL in an expanded intermediate state. *Nature* **461,** 120–124 (2009).
36. Bass, R. B., Strop, P., Barclay, M. & Rees, D. C. Crystal structure of *Escherichia coli* MscS, a voltage-modulated and mechanosensitive channel. *Science* **298,** 1582–1587 (2002).
37. Chang, G., Spencer, R. H., Lee, A. T., Barclay, M. T. & Rees, D. C. Structure of the MscL homolog from *Mycobacterium tuberculosis*: a gated mechanosensitive ion channel. *Science* **282,** 2220–2226 (1998).
38. Kung, C., Martinac, B. & Sukharev, S. Mechanosensitive channels in microbes. *Annu. Rev. Microbiol.* **64,** 313–329 (2010).
39. Brohawn, S. G., del Mármol, J. & MacKinnon, R. Crystal structure of the human K2P TRAAK, a lipid- and mechano-sensitive K^+ ion channel. *Science* **335,** 436–441 (2012).
40. Coste, B. *et al.* Piezo1 ion channel pore properties are dictated by C-terminal region. *Nature Commun.* **6,** 7223 (2015).
41. Hou, X., Pedi, L., Diver, M. M. & Long, S. B. Crystal structure of the calcium release-activated calcium channel Orai. *Science* **338,** 1308–1313 (2012).
42. Penna, A. *et al.* The CRAC channel consists of a tetramer formed by Stim-induced dimerization of Orai dimers. *Nature* **456,** 116–120 (2008).
43. Kamajaya, A., Kaiser, J. T., Lee, J., Reid, M. & Rees, D. C. The structure of a conserved Piezo channel domain reveals a topologically distinct β sandwich fold. *Structure* **22,** 1520–1527 (2014).
44. Prole, D. L. & Taylor, C. W. Identification and analysis of putative homologues of mechanosensitive channels in pathogenic protozoa. *PLoS One* **8,** e66068 (2013).
45. Gonzales, E. B., Kawate, T. & Gouaux, E. Pore architecture and ion sites in acid-sensing ion channels and P2X receptors. *Nature* **460,** 599–604 (2009).
46. Vilceanu, D. & Stucky, C. L. TRPA1 mediates mechanical currents in the plasma membrane of mouse sensory neurons. *PLoS One* **5,** e12177 (2010).
47. Kwan, K. Y., Glazer, J. M., Corey, D. P., Rice, F. L. & Stucky, C. L. TRPA1 modulates mechanotransduction in cutaneous sensory neurons. *J. Neurosci.* **29,** 4808–4819 (2009).
48. Smart, O. S., Goodfellow, J. M. & Wallace, B. A. The pore dimensions of gramicidin A. *Biophys. J.* **65,** 2455–2460 (1993).

Acknowledgements We thank H. Yu and J. Chai for discussion and proofreading of the manuscript. We thank the staff at beamline BL17U of the Shanghai Synchrotron Radiation Facility (SSRF) and beamline 3W1A of the Beijing Synchrotron Radiation Facility (BSRF) for their assistance in data collection. K. Wu and H. Wang are acknowledged for technique help. We also thank the National Center for Protein Sciences (Beijing, China) for technical support with cryo-electron microscopy data collection and for computation resources. This work was supported by grants from the Ministry of Science and Technology (2012CB911101 and 2011CB910502 to M.Y., 2015CB910102 to B.X. and 2013CB910404 to N.G.), the National Natural Science Foundation of China (21532004, 31570733, 31030020 and 31170679 to M.Y., 31422016 to N.G. and 31422027 to B.X.) and the Ministry of Education (the Young Thousand Talent program to B.X.).

Author Contributions M.Y. directed the study. J.G., M.C. and R.L. performed protein purification, detergent screening and crystallization. W.L. performed electron microscopy sample preparation, data collection and structural determination with N.L; Q.Z. was responsible for molecular cloning (with P.Z.), protein purification, detergent screening and biochemical and confocal imaging studies. N.G. directed electron microscopy studies and wrote part of the manuscript. B.X. initiated the project and directed molecular cloning, protein expression and purification and wrote most of the manuscript. All authors contributed to discussion of the data and editing of the manuscript.

Author Information The 3D cryo-electron microscopy density map has been deposited in the Electron Microscopy Data Bank (EMDB), with accession code EMD-6343. The coordinates of atomic models have been deposited in the Protein Data Bank (PDB) under the accession codes 4RAX for the CED and 3JAC for the full length. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to N.G. (ninggao@tsinghua.edu.cn), B.X. (xbailong@biomed.tsinghua.edu.cn) and M.Y. (maojunyang@tsinghua.edu.cn).

METHODS

No statistical methods were used to predetermine sample size, the experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

Molecular cloning. The pcDNA3.1-Piezo1-pp (PPase, PreScission protease cleavage site) -GST-IRES-GFP construct was subcloned by inserting the coding sequence of the PreScission protease cleavage site between Piezo1 (E2JF22, UniprotKB entry) and GST coding sequences in the parental construct of pcDNA3.1-Piezo1-GST-IRES-GFP[3]. Piezo1-Cterm-Flag-IRES-GFP was subcloned by inserting the synthesized double-stranded DNA fragment encoding Flag between the Piezo1-coding sequence and IRES-GFP using the restriction enzymes AscI and SacII. Piezo1(A2419)-Flag-IRES-GFP was constructed using a one-step cloning kit (Vazyme Biotech) by introducing the Flag-tag coding sequence after the residue Piezo1(A2419) into the Piezo1-GST-IRES-GFP construct and the Piezo1(Δ2219–2453) construct was generated by deleting amino acids 2219–2453 from the Piezo1-pp-GST-IRES-GFP construct. The coding sequence of the CED of Piezo1 (residues 2214–2457) was cloned into a pET22b (Novagen) vector with a C-terminal 6×His tag using the restriction enzymes NdeI and XhoI. All the constructs were confirmed by sequencing.

Protein expression and purification of Piezo1 and Piezo1(Δ2219–2453). HEK293T cells were grown in DMEM (basic) with 10% FBS. When the density of cells cultured in 150 mm × 25 mm dishes reached 80–90%, the expression plasmids were transiently transfected with polyethylenimines (Polysciences). The protein purification procedure was slightly modified from similar previously described methods[3]. After 48 h, the transfected cells were collected, washed twice with PBS and homogenized in buffer A, containing 25 mM Na-PIPES, pH 7.2, 140 mM NaCl, 2 mM dithiothreitol (DTT), detergents CHAPS (1%) and C12E9 (0.1%), 0.5% (w/v) L-α-phosphatidylcholine (Avanti) and a cocktail of protease inhibitors (Roche) at 4 °C for 1 h. After centrifugation at 100,000*g* for 40 min, the supernatant was collected and incubated with glutathione–sepharose beads (GE Healthcare) at 4 °C for 3 h. The resin was washed extensively with buffer B, containing 25 mM Na-PIPES, pH 7.2, 140 mM NaCl, 2 mM DTT, 0.1% (w/v) C12E9 and 0.01% (w/v) L-α-phosphatidylcholine. The GST-free or GST-tagged Piezo1 was cleaved off by PreScission Protease (Amersham-GE) in buffer B at 4 °C overnight or directly eluted from the protein-loaded resin with buffer B plus 10 mM GSH, respectively, and applied to size-exclusion chromatography (Superpose-6 10/300 GL, GE Healthcare) in buffer C (25 mM Na-PIPES, pH 7.2, 140 mM NaCl, 2 mM DTT) plus 0.026% (w/v) C12E10 or other detergents in the final concentration of 2× critical micelle concentration. For amphipol-bound Piezo1, amphipols were substituted for detergents as described[34], after which the protein was loaded on a Superpose-6 column in buffer C. Proteins with different kinds of detergents or amphipols were examined by both gel filtration and negative staining. Peak fractions representing oligomeric Piezo1 were collected for electron microscopy analysis. Protein in C12E10 was used for final cryo-electron microscopy structure determination. All detergents and amphipols used in this project were purchased from Anatrace.

Expression and purification of Piezo1 CED fragment. Overexpression of Piezo1 CED was induced in *Escherichia coli* BL21 strain by 0.5 mM isopropyl-β-D-thiogalactoside when the cell density reached an optical density of ~0.8 at 600 nm. After growing at 18 °C for 12 h, the cells were collected, washed, resuspended in buffer D, containing 25 mM Tris-HCl, pH 8.0, 500 mM NaCl and 20 mM imidazole, and lysed by sonication. The lysates were clarified by centrifugation at 23,000*g* for 1 h and the supernatant was collected and loaded onto Ni^{2+}-nitrilotriacetate affinity resin (Ni-NTA, Qiagen). The resin was washed extensively with buffer D and eluted with buffer D plus 280 mM imidazole. The eluate was concentrated and subjected to gel filtration (Superdex-200, GE Healthcare) with buffer E, containing 25 mM Tris-HCl, pH 8.0, 200 mM NaCl, 2 mM DTT, or buffer F, containing 25 mM Tris-HCl, pH 8.0, 25 mM NaCl and 2 mM DTT (Extended Data Fig. 6e).

NativePAGE Novex Bis-Tris gel and western blotting. The purified Piezo1 proteins were subjected to 3–12% NativePAGE Novex Bis-Tris gel for native electrophoresis according to the manufacturer's protocol at 150 V for 2 h. The native gel was transferred to a positively charged nylon/nitrocellulose membrane at 100 V for 1.5 h. After incubating in 8% (v/v) acetic acid to fix the proteins, air drying and rewetting with methanol, the membrane was blocked with 5% (w/v) milk in TBS buffer with 0.1% (w/v) Tween-20 (TBST buffer) at room temperature (~26 °C) for 1 h. The membrane was then incubated with the anti-Piezo1 antibody (1:1,000) (custom generated using the peptide YIRAPNGPEANPVK) at room temperature for 1 h, followed by washing with TBST buffer and further incubated with anti-rabbit IgG antibody (1:10,000) at room temperature for 1 h. Proteins were detected with the SuperSignal West Pico Chemiluminescent Substrate (Thermo).

Immunostaining. For live-cell labelling, cells grown on coverslips were incubated with the anti-Flag antibody (1:100, Sigma) diluted in prewarmed culture medium at room temperature for 1 h. After three washes, cells were incubated with the Alexa Fluor 594 donkey-anti-mouse IgG secondary antibody (1:200, Life Technologies) at room temperature for 1 h and then washed and fixed with 4% (w/v) paraformaldehyde. For permeabilized staining, cells were first fixed with 4% (w/v) paraformaldehyde and permeabilized with 0.2% (w/v) Triton X-100, then incubated with the anti-Flag antibody (1:200, Sigma) or the anti-GST antibody (1:200, Millipore) at room temperature for 1 h. Cells were washed and then incubated with the Alexa Fluor 594 donkey-anti-mouse IgG (1:200, Life Technologies) or Alexa Fluor 594 donkey-anti-rabbit IgG (1:200, Life Technologies) secondary antibody at room temperature for 1 h. After washing, coverslips were mounted and imaged using a Nikon A1 confocal microscope with a 60× oil objective (N.A. = 1.49) at either the GFP (488-nm exciting wavelength) or the TRITC channel (561-nm exciting wavelength).

Crystallization, data collection and structure determination of the CED. Crystals of CED proteins were obtained at 18 °C using the sitting-drop method by mixing 1 μl protein (15 mg ml^{-1}) with 1 μl reservoir solution (0.1 M HEPES, pH 7.5, 0.2 M $MgCl_2$ and 25% w/v PEG3350). Crystals appeared after 2–3 weeks and reached full size in about a month. The crystals were cryo-protected in reservoir solution containing 15–20% glycerol and flash frozen in liquid nitrogen before data collection. Native data of CED crystals were collected at beamline BL17U of the Shanghai Synchrotron Radiation Facility (SSRF). Single-wavelength anomalous dispersion data were collected at 100 K using a MAResearch M165 charge-coupled device (CCD) detector at the Beijing Synchrotron Radiation Facility (BSRF), with the crystals soaked in 2 M NaI for 1 min. All diffraction data were processed with HKL2000 (ref. 49). Further processing was carried out using programs from the CCP4 suite (Collaborative Computational Project)[50]. The heavy-atom positions in the iodine-soaked crystal were determined using SHELXD[51]. Heavy-atom parameters were then refined and initial phases were generated in the program PHASER[52] using the single-wavelength anomalous dispersion experimental phasing module. The real-space constraints were applied to the electron density map in DM[53]. The resulting map was of sufficient quality for building the model of the CED in Coot[54]. The structures were refined with the PHENIX packages[55]. Full data collection and structure statistics are summarized in Extended Data Table 1.

Negative-staining electron microscopy. An aliquot of 4 μl Piezo1 (0.05 mg ml^{-1}) was applied to glow-discharged carbon-coated copper grids (200 mesh, Zhongjingkeyi, Beijing). After the grids were incubated at room temperature for 1 min, excessive liquid was absorbed by filter paper. Grids containing the specimen were stained by applying droplets of 2% uranyl acetate for 30 s and air dried. Micrographs were generated on a T12 microscope (FEI) operated at 120 kV, using a 4k × 4k CCD camera (UltraScan 4000, Gatan). Images of Piezo1 purified with C12E10, C12E8 and amphipol A8-35 were recorded at a nominal magnification of 68,000× and with a pixel size of 1.59 Å (Extended Data Fig. 2). Images of Piezo1(ΔCED) in C12E10 were recorded at a nominal magnification of 49,000× and with a pixel size of 2.21 Å. Micrographs of random conical tilt (RCT) pairs were taken at 50° and 0° tilt angles at a nominal magnification of 49,000×.

Cryo-electron microscopy. The detergent C12E10 was chosen for cryo-electron microscopy analysis because it produced slightly better micrographs (Extended Data Fig. 2). Aliquots of 4 μl detergent-solubilized (C12E10) Piezo1 at a concentration of 0.2 mg ml^{-1} were applied to glow-discharged 300-mesh Quantifoil R2/2 grids (Quantifoil, Micro Tools GmbH, Germany) coated with a self-made continuous thin carbon. After 15 s of waiting time, grids were blotted for 1.5 s and plunged into liquid ethane using an FEI Mark IV Vitrobot operated at 4 °C and 100% humidity. Grids were examined using a TF20 microscope (FEI) operated at 200 kV with a nominal magnification of 62,000× and images were captured on a CCD camera (Gatan) under low-dose conditions. High-resolution images were captured on a Titan Krios microscope, operated at 300 kV, with a K2 Summit direct electron detector (Gatan) in counting mode. Data acquisition was performed using UCSF-Image4 (X. Li and Y. Cheng), with a nominal magnification of 22,500×, which yields a final pixel size of 1.32 Å at object scale and with defocus ranging from −1.7 μm to −2.9 μm. The dose rate on the detector was about 8.2 counts per pixel per second, with a total exposure time of 8 s. Each micrograph stack consists of 32 frames.

Image processing. The data sets of negative-staining electron microscopy were processed with EMAN2.1 (ref. 56) and RELION[57]. Reference-free 2D classification was performed with RELION. The numbers of Piezo1 particles in the presence of C12E10, C12E8 and amphipol A8-35 are 7,279, 14,045 and 7,565, respectively. For RCT[58] data processing, particle picking and classification were performed with EMAN2.1 (ref. 56) and reconstruction of RCT classes and structural refinement from all untilted particles were performed with SPIDER[59]. The

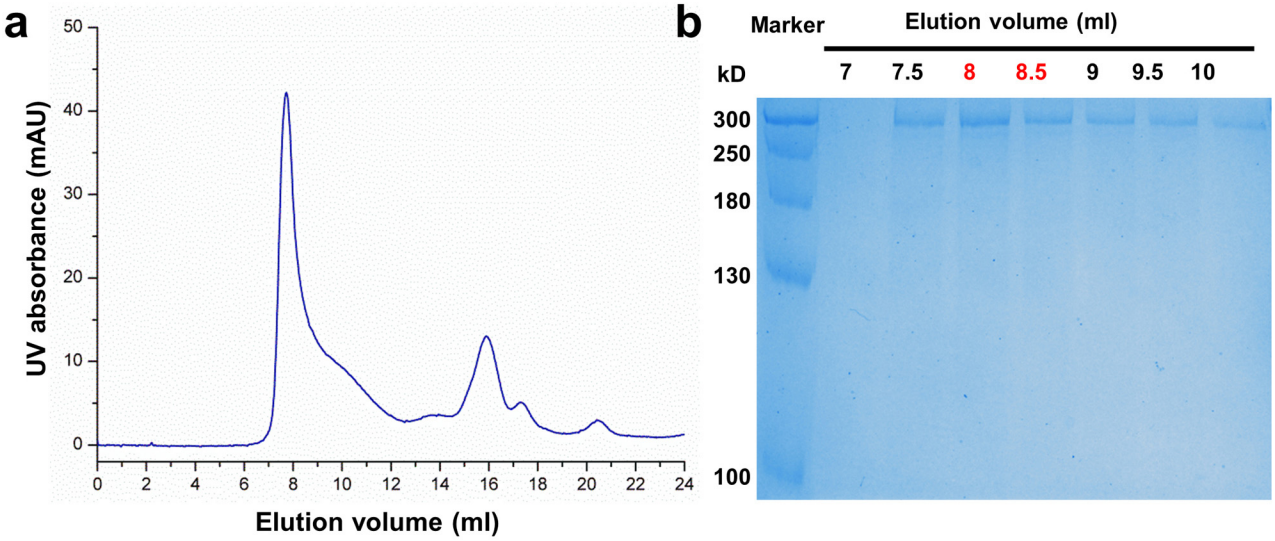
final number of particles used in generating the initial model is 5,670. The initial 3D reference created using the RCT method is shown in Extended Data Fig. 3.

For cryo-electron microscopy (TF20) data processing, 505 micrographs were processed with SPIDER[59] and RELION[57]. Particles were picked using SPIDER, manually screened (39,555 in total) and subjected to reference-free 2D classification using RELION. A final number of 16,729 particles were used for 3D refinement using the RCT model as initial reference. To validate the 3D model, 3D refinement was also performed with a Gaussian density ball as initial reference. During refinement, both the symmetry-free (C1) and symmetry-imposed (C3) reconstructions were tested (Extended Data Fig. 3d).

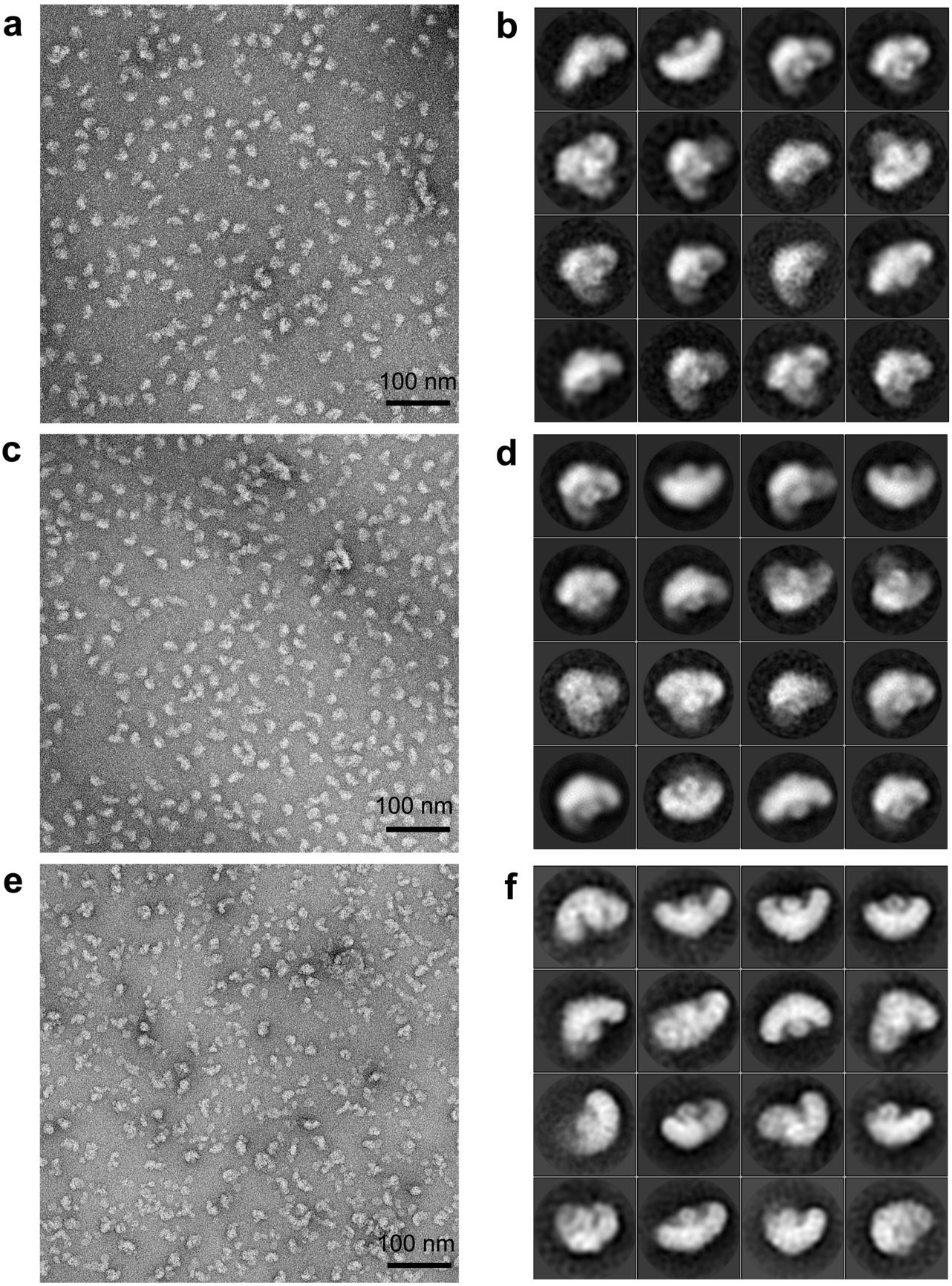
For processing K2 micrographs, motion correction was applied at the micrograph level using the dosefgpu_driftcorr program (developed by X. Li) to produce average micrographs across all frames[60]. Micrograph screening, particle picking and normalization were performed with SPIDER. The program CTFFIND3 (ref. 61) was used to estimate the contrast transfer function parameters. The 2D and 3D classification and refinement were performed with RELION exclusively to avoid potential structural overfitting. Classification of raw cryo-electron microscopy particles resulted in well-resolved 2D class averages, with many secondary structural features clearly discernable. In particular, on class averages of typical side views, many pieces of rod-like densities arranged in parallel fashion could be readily identified, raising the possibility that they were transmembrane helices (Fig. 2c). A total of 179,805 particles from 1,042 micrographs were subject to a cascade of 2D and 3D classification (Extended Data Fig. 5a). During 3D classification, no symmetry was imposed. Different combinations of particles from these classes were tested in refinement. After two rounds of 3D classification, a set of adequately homogeneous particles (30,021), which best matched the C3 symmetry, was subjected to a third round of 3D classification. This resulted in generally similar class structures, with no detectable improvement on particle homogeneity. Consequently, this set of particles was used for final refinement, with the RCT model low-pass filtered to 60 Å as initial reference. Applying the C3 symmetry in the refinement resulted in an overall structure at a resolution of 10.24 Å. After the first refinement, we noted that translation parameters of particles (OriginX and OriginY in RELION) were rather large, with many particles having x or y shifts of more than 15 pixels. Particles were rewindowed from original micrographs by applying their x and y shifts. Rewindowed particles were subjected to a second round of refinement using RELION, which only marginally improved the density map. A third round of refinement was performed by applying an enlarged soft mask (Extended Data Fig. 5a) of the Piezo1 channel, which improved the overall resolution to 6.03 Å. Last, particle-based beam-induced movement correction was performed by statistical movie processing in RELION, using movie frames 2–15. This yielded a final 3D density map with an overall resolution of 5.9 Å, with regions defined by the soft mask being 4.8 Å (Extended Data Fig. 5b). All reported resolutions are based on the gold-standard FSC = 0.143 (ref. 62) and the final FSC curve (4.8 Å) was corrected for the effect of a soft mask using high-resolution noise substitution[63]. In addition, subregion refinements, as previously described for ribosomal complex structural determination[64–67], were applied to improve the local densities of interest, by using a soft mask of the cap domain, the lower central pore region and a single subunit. The subsequent reported resolutions were still in the range of 4.8–5.5 Å, but with much-improved densities for these masked regions. This led to a separation of secondary structural elements in the cap and transmembrane regions. However, in all cases, the densities at the distal 'blade' domain are fragmented and limited our further quantitative analysis. Final density maps were sharpened by a B-factor of –100 Å^2 using RELION. A local resolution map was calculated using ResMap[68]. UCSF Chimera[69] was used to fit the crystal structure of the CED to the density map of the cap domain.

Poly-alanine model and structural analysis. Main-chain tracing and building a poly-alanine model were done manually using Coot[70]. Sequence alignment was performed using Clustal W2 (ref. 71). Secondary structures were predicted with PredictProtein[72] using the full-length Piezo1 sequence. Transmembrane segments were predicted using multiple prediction web servers, including Topcons[73], TMHMM2 (ref. 74), HMMTOP[75] and Phobius[76], with their results shown as green, blue, orange and pink lines, respectively, in Extended Data Fig. 9. Sequence alignment and secondary structure prediction of Piezo1 from different species were used to aid the assignment of structural elements in the density map. Multiple rounds of model rebuilding in Coot were performed for model optimization.

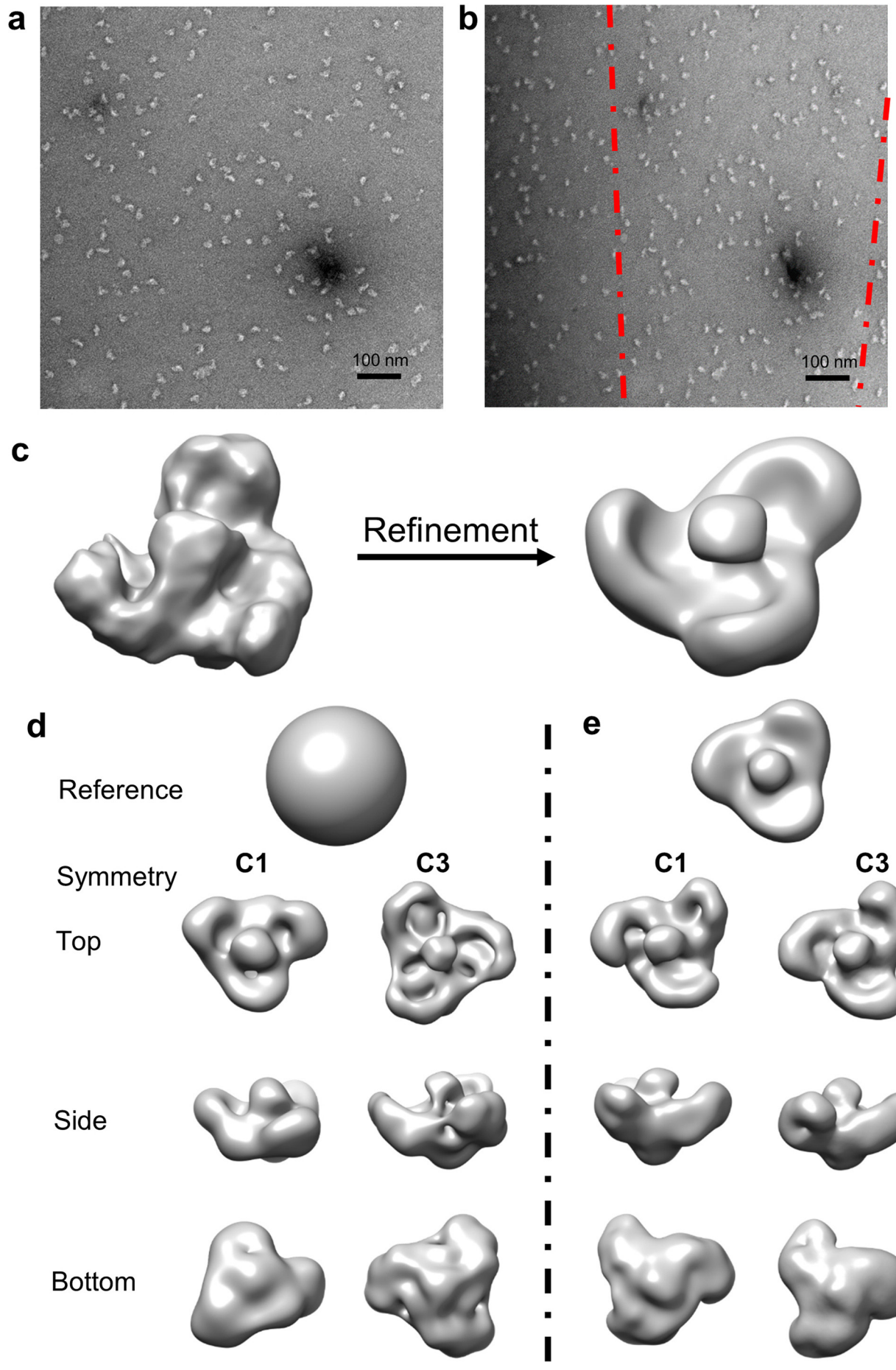
49. Otwinowski, Z. & Minor, W. Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol.* **276,** 307–326 (1997).
50. Winn, M. D. *et al.* Overview of the CCP4 suite and current developments. *Acta Crystallogr. D* **67,** 235–242 (2011).
51. Schneider, T. R. & Sheldrick, G. M. Substructure solution with SHELXD. *Acta Crystallogr. D* **58,** 1772–1779 (2002).
52. McCoy, A. J. *et al.* Phaser crystallographic software. *J. Appl. Crystallogr.* **40,** 658–674 (2007).
53. Cowtan, K. D. & Main, P. Improvement of macromolecular electron-density maps by the simultaneous application of real and reciprocal space constraints. *Acta Crystallogr. D* **49,** 148–157 (1993).
54. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr. D* **60,** 2126–2132 (2004).
55. Adams, P. D. *et al.* PHENIX: building new software for automated crystallographic structure determination. *Acta Crystallogr. D* **58,** 1948–1954 (2002).
56. Tang, G. *et al.* EMAN2: an extensible image processing suite for electron microscopy. *J. Struct. Biol.* **157,** 38–46 (2007).
57. Scheres, S. H. RELION: implementation of a Bayesian approach to cryo-EM structure determination. *J. Struct. Biol.* **180,** 519–530 (2012).
58. Radermacher, M., Wagenknecht, T., Verschoor, A. & Frank, J. Three-dimensional reconstruction from a single-exposure, random conical tilt series applied to the 50S ribosomal subunit of *Escherichia coli*. *J. Microsc.* **146,** 113–136 (1987).
59. Shaikh, T. R. *et al.* SPIDER image processing for single-particle reconstruction of biological macromolecules from electron micrographs. *Nature Protocols* **3,** 1941–1974 (2008).
60. Li, X. *et al.* Electron counting and beam-induced motion correction enable near-atomic-resolution single-particle cryo-EM. *Nature Methods* **10,** 584–590 (2013).
61. Mindell, J. A. & Grigorieff, N. Accurate determination of local defocus and specimen tilt in electron microscopy. *J. Struct. Biol.* **142,** 334–347 (2003).
62. Scheres, S. H. & Chen, S. Prevention of overfitting in cryo-EM structure determination. *Nature Methods* **9,** 853–854 (2012).
63. Chen, S. *et al.* High-resolution noise substitution to measure overfitting and validate resolution in 3D structure determination by single particle electron cryomicroscopy. *Ultramicroscopy* **135,** 24–35 (2013).
64. Voorhees, R. M., Fernández, I. S., Scheres, S. H. & Hegde, R. S. Structure of the mammalian ribosome-Sec61 complex to 3.4 Å resolution. *Cell* **157,** 1632–1643 (2014).
65. Greber, B. J. *et al.* The complete structure of the large subunit of the mammalian mitochondrial ribosome. *Nature* **515,** 283–286 10.1038/nature13895 (2014).
66. Brown, A. *et al.* Structure of the large ribosomal subunit from human mitochondria. *Science* **346,** 718–722 (2014).
67. Fernández, I. S., Bai, X. C., Murshudov, G., Scheres, S. H. & Ramakrishnan, V. Initiation of translation by cricket paralysis virus IRES requires its translocation in the ribosome. *Cell* **157,** 823–831 (2014).
68. Kucukelbir, A., Sigworth, F. J. & Tagare, H. D. Quantifying the local resolution of cryo-EM density maps. *Nature Methods* **11,** 63–65 (2014).
69. Pettersen, E. F. *et al.* UCSF Chimera – a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25,** 1605–1612 (2004).
70. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. D* **66,** 486–501 (2010).
71. Larkin, M. A. *et al.* Clustal W and Clustal X version 2.0. *Bioinformatics* **23,** 2947–2948 (2007).
72. Yachdav, G. *et al.* PredictProtein – an open resource for online prediction of protein structural and functional features. *Nucleic Acids Res.* **42,** W337–W343 (2014).
73. Bernsel, A., Viklund, H., Hennerdal, A. & Elofsson, A. TOPCONS: consensus prediction of membrane protein topology. *Nucleic Acids Res.* **37,** W465–W468 (2009).
74. Sonnhammer, E. L., von Heijne, G. & Krogh, A. A hidden Markov model for predicting transmembrane helices in protein sequences. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **6,** 175–182 (1998).
75. Tusnády, G. E. & Simon, I. The HMMTOP transmembrane topology prediction server. *Bioinformatics* **17,** 849–850 (2001).
76. Käll, L., Krogh, A. & Sonnhammer, E. L. Advantages of combined transmembrane topology and signal peptide prediction – the Phobius web server. *Nucleic Acids Res.* **35,** W429–W432 (2007).



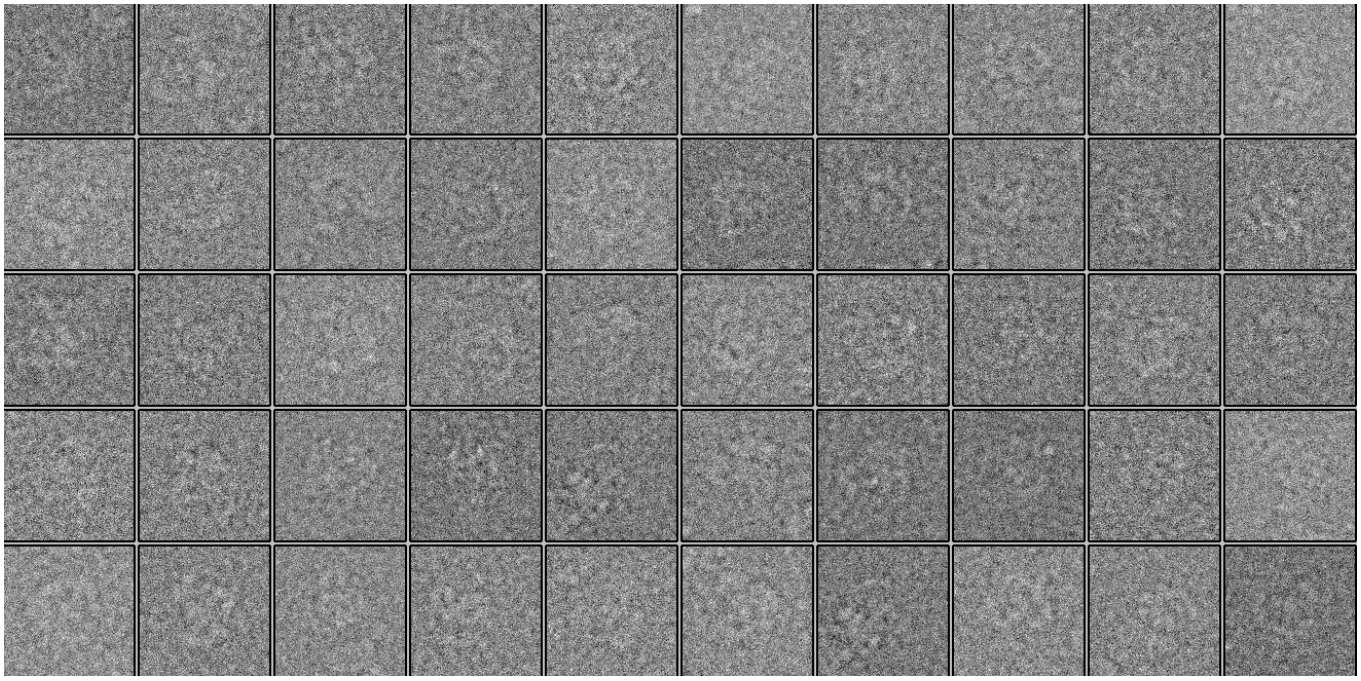
Extended Data Figure 1 | Biochemical characterization of the recombinant protein of Piezo1–pp–GST. **a**, A representative trace of gel filtration chromatography of the Piezo1–pp–GST protein. **b**, Protein samples of the indicated fractions were subjected to SDS–PAGE and Coomassie blue staining. Fractions of 8.0 ml and 8.5 ml (elution volume) were used for the negative-staining electron microscopy and native gel analyses, respectively.



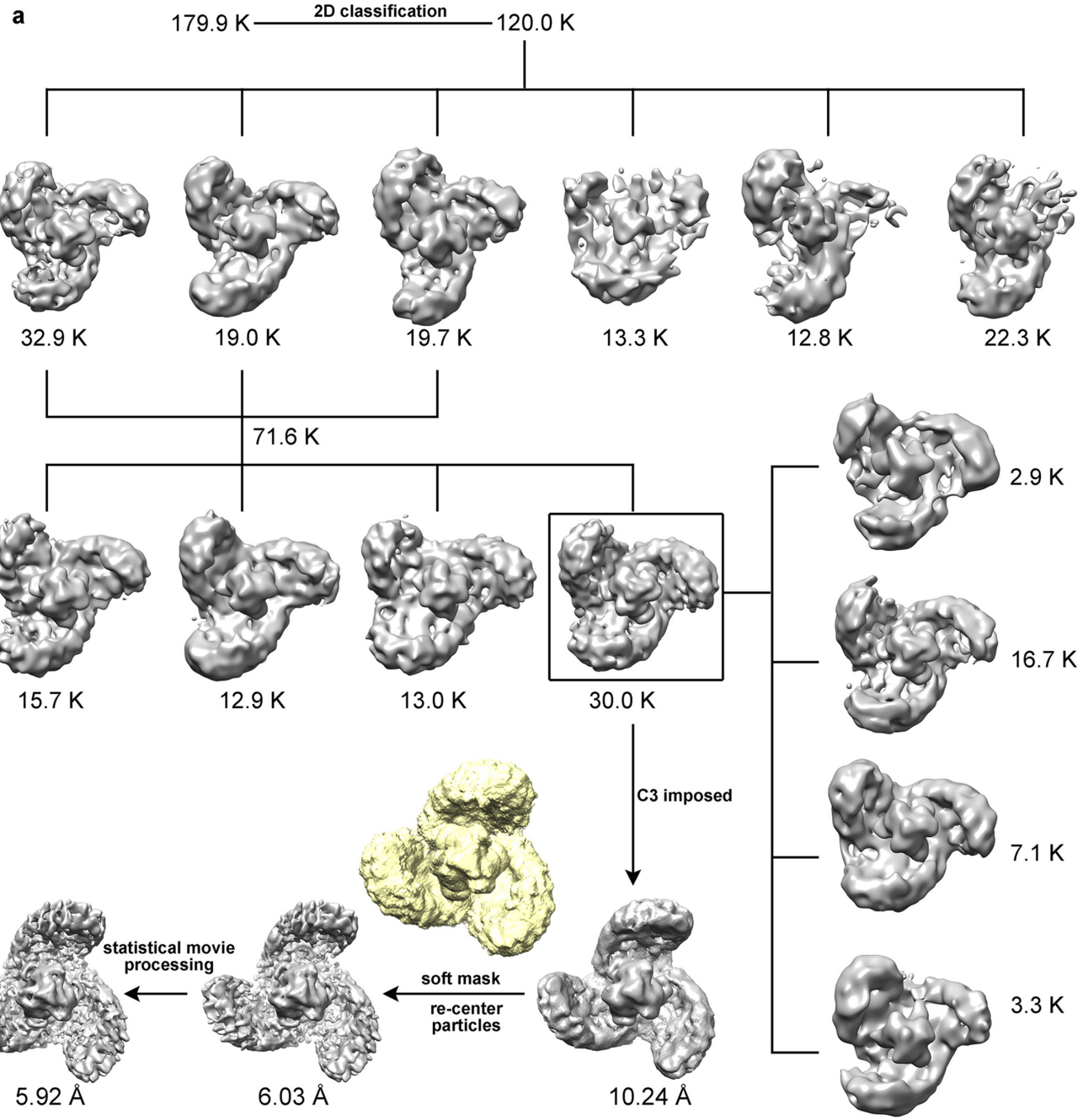
Extended Data Figure 2 | Negative-staining electron microscopy examination of Piezo1 in different detergents. **a**, A representative micrograph of negatively stained Piezo1 purified with C12E10. **b**, 2D class averages of Piezo1 particles (C12E10). **c**, A representative micrograph of negatively stained Piezo1 purified with C12E8. **d**, 2D class averages of Piezo1 particles (C12E8). **e**, A representative micrograph of negatively stained Piezo1, with amphipol A8-35 as detergent. **f**, 2D class averages of Piezo1 particles (amphipol A8-35).

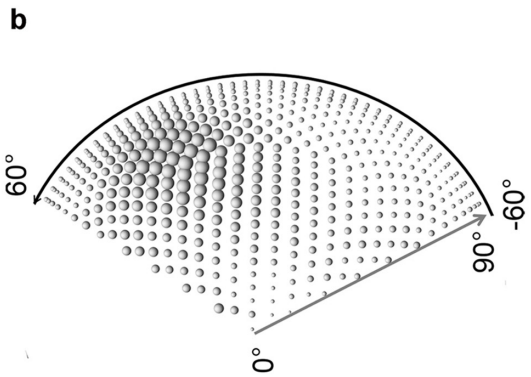


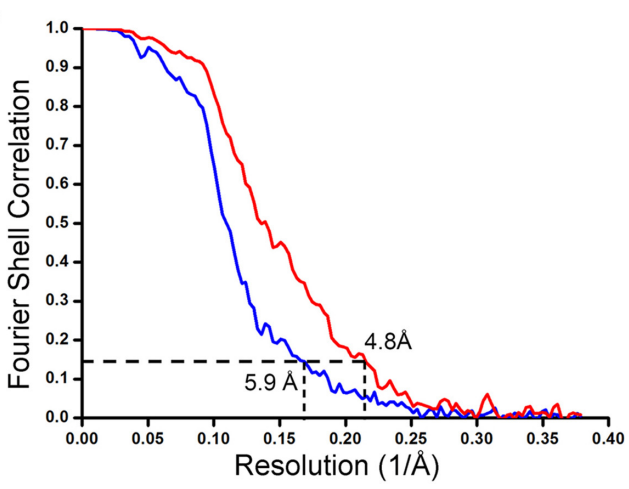
Extended Data Figure 3 | Initial model of Piezo1 generated from the random conical tilt method and validation of the model using cryo-electron microscopy data from a TF20 microscope. **a**, **b**, Representative micrographs of negatively stained Piezo1 in C12E10 collected in random conical tilt (RCT) pairs (**a**, untilted and **b**, 50° tilted). **c**, Top view of an RCT reconstruction, showing an overall threefold symmetry for the Piezo1 complex, is shown on the left. The right-hand side shows the top view of the refined model, obtained by a structural refinement of all particles from untilted micrographs. **d**, **e**, Model validation was performed by refinement of cryo-electron microscopy particles collected with TF20, with a Gaussian ball (**d**) or the RCT model (**e**) as initial reference. The 3D volumes are shown in top, side and bottom views. During the refinement, both the symmetry-free (C1) and symmetry-imposed (C3) reconstructions were tested. Note that some of these reconstructions have incorrect handedness.



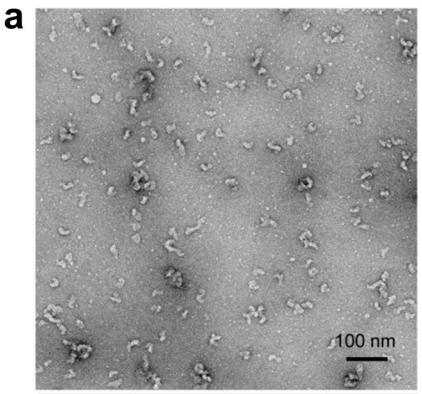
Extended Data Figure 4 | Representative raw particles of Piezo1 collected with the Titan Krios electron microscope fitted with a K2 electron detector. A collection of raw particles of Piezo1 (eluted with C12E10), collected with Titan Krios (300 kV).

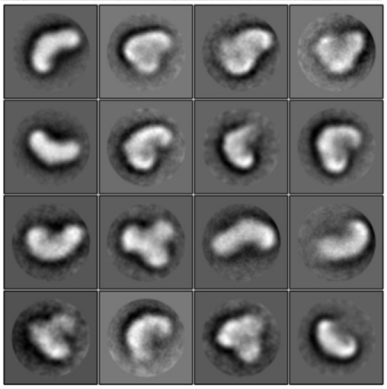


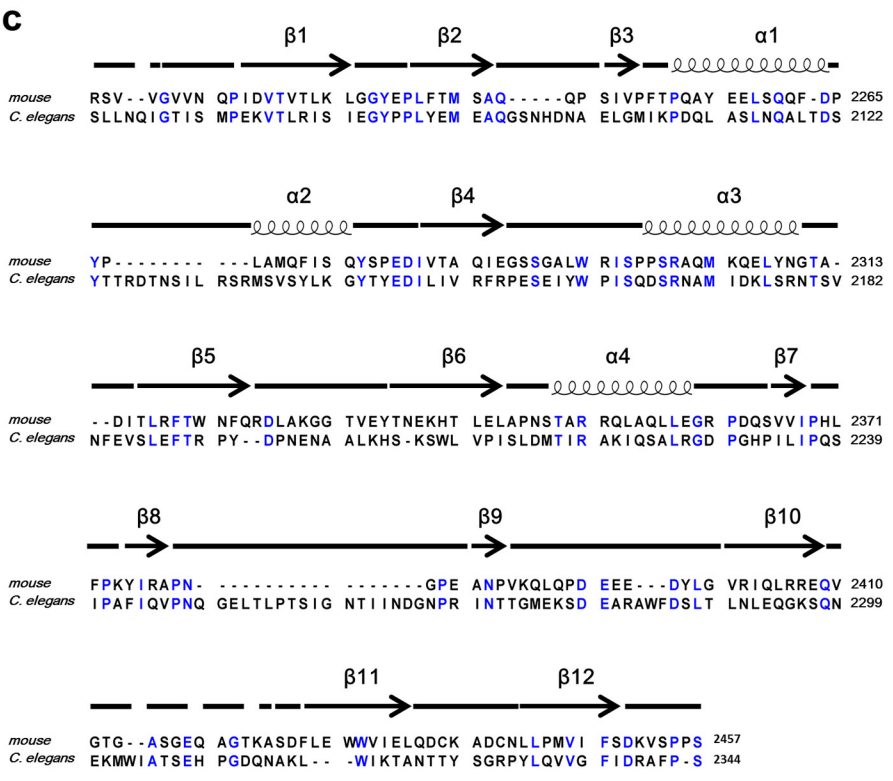


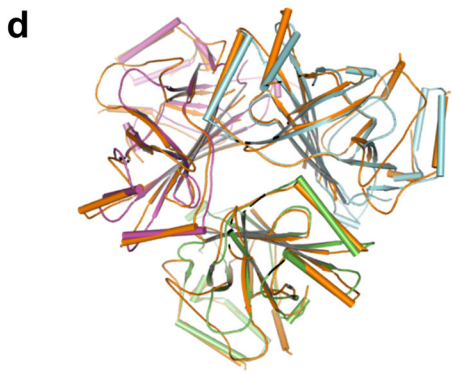


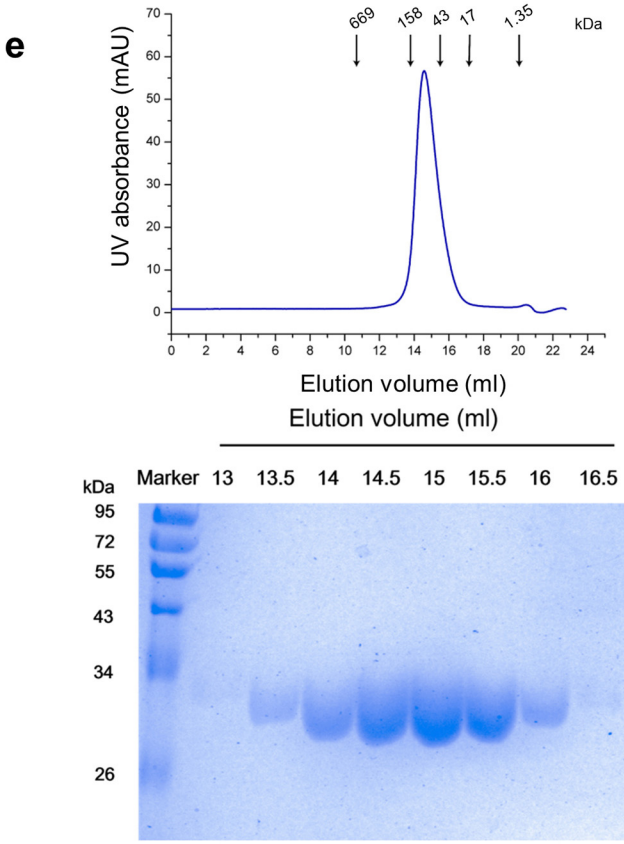
Extended Data Figure 5 | Workflow of 3D classification of Piezo1 particles. **a**, The schematic diagram of a series of 3D classification procedures with RELION is shown (also see Methods). After several rounds of 2D classification, the remaining 120,000 particles were subjected to three rounds of 3D classification without imposing any symmetry. A final set of particles (class 4 after the second round of 3D classification), with its reconstruction best matching threefold symmetry, was subjected to 3D refinement (C3 imposed). Notably, further 3D classification of this class resulted in generally similar structures (vertically arranged panels) without detectable improvement of conformational homogeneity. A top view of the soft mask used in structural refinement is also shown (yellow). **b**, Distribution of particle orientations in the last iteration of the refinement. **c**, Gold-standard Fourier shell correlation (FSC) curves of the final density map. The FSC curves were calculated with (red) or without (blue) the application of a soft mask to the two half-set maps. The final FSC curve (red) was corrected for the soft-mask-induced effect. Reported resolutions were based on FSC = 0.143 criteria.

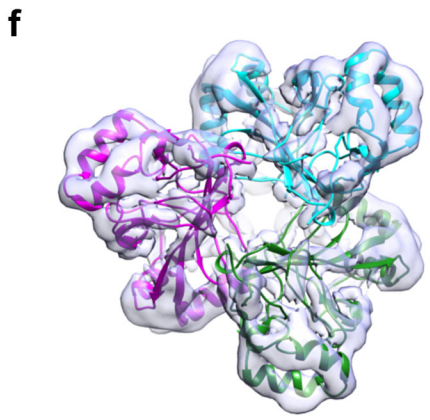




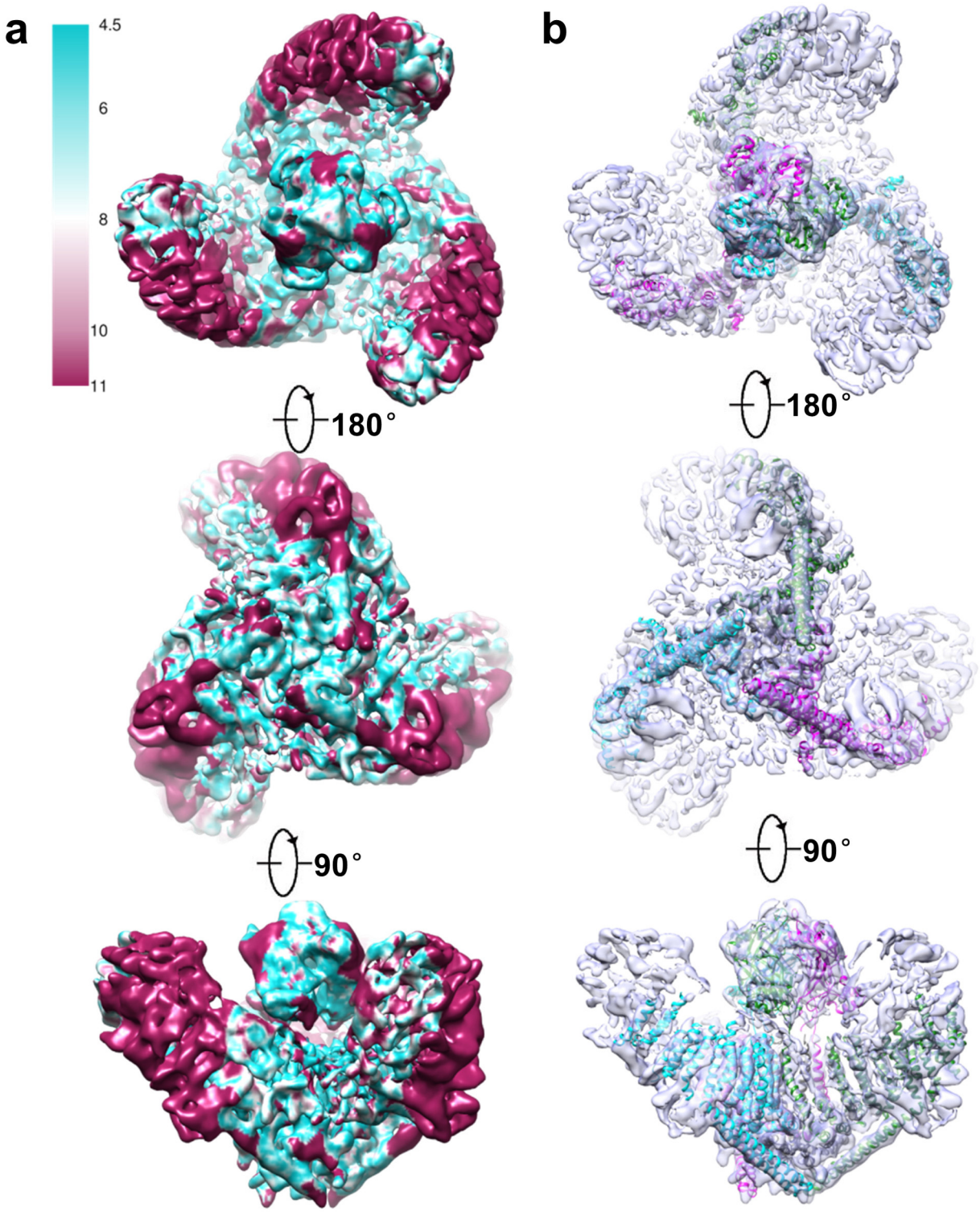




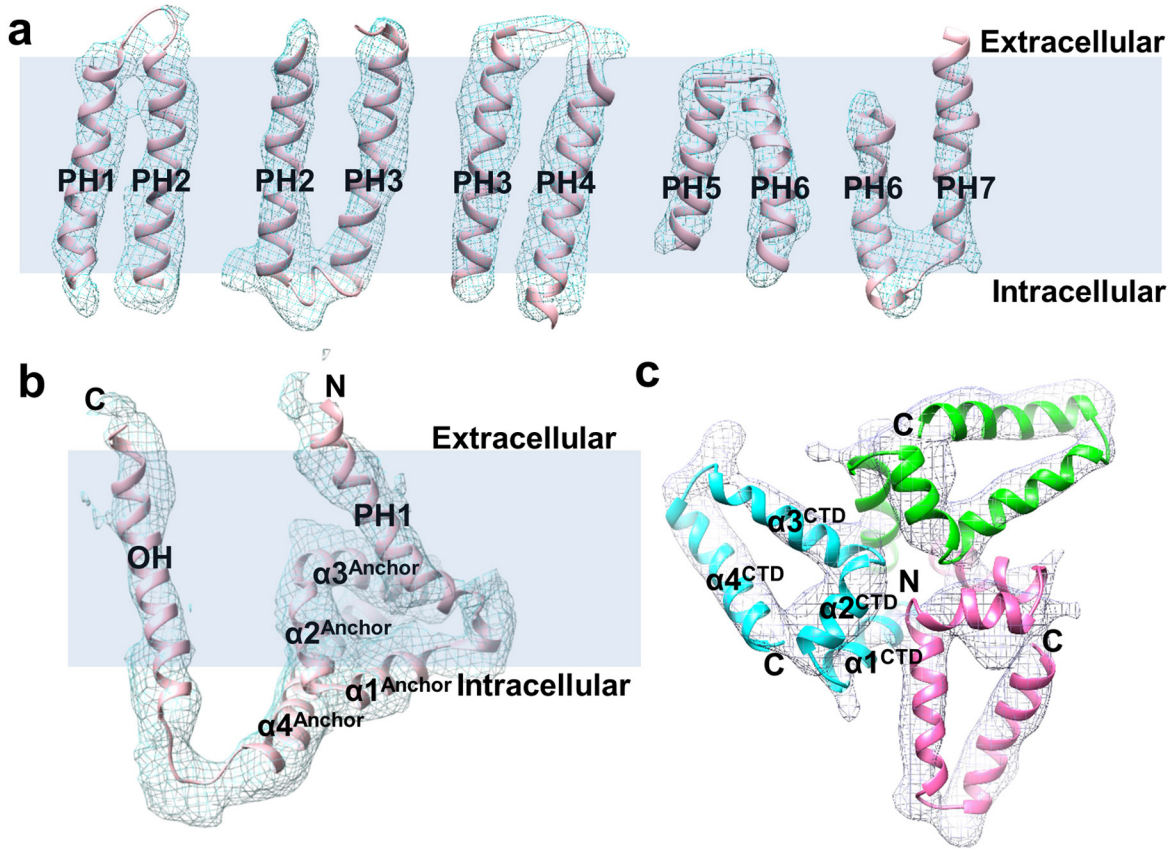




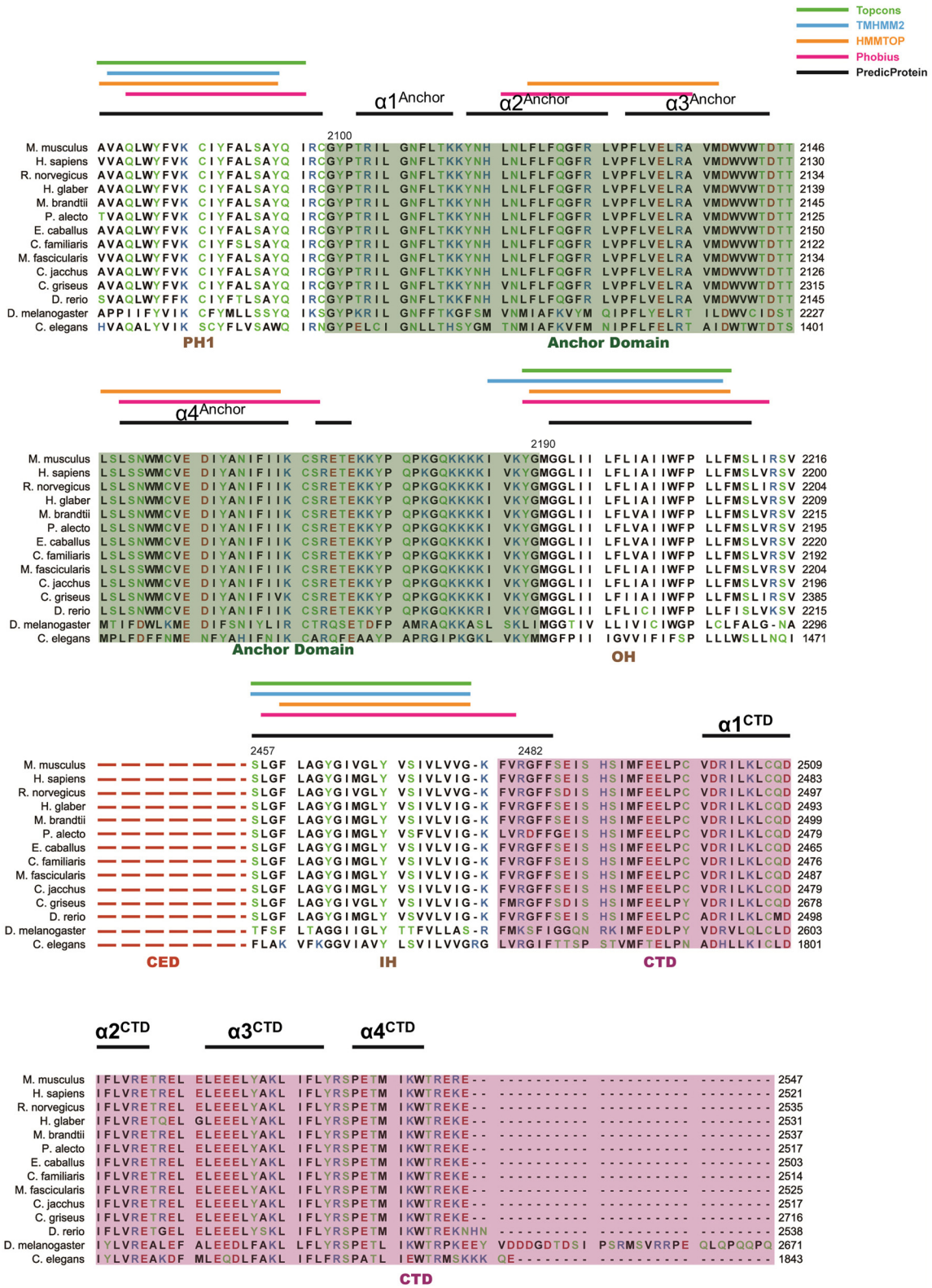
Extended Data Figure 6 | The trimeric CEDs form the cap domain of Piezo1. **a**, A representative micrograph of negatively stained Piezo1(ΔCED) in C12E10. **b**, 2D class averages of negatively stained Piezo1(ΔCED) particles. It is evident that the central cap domain is absent from these average images. **c**, Sequence alignment of the CED region of Piezo1 from *Mus musculus* and *Caenorhabditis elegans*. Identical residues are highlighted in blue. Secondary structures are indicated by cartoons above the primary sequence. Sequence alignment was performed using Clustal W2 (ref. 71). **d**, Structure alignment of the trimeric CED of Piezo1 from *M. musculus* and *C. elegans*. The three CEDs are coloured in purple, cyan and green, respectively. The CED of *C. elegans* is coloured in orange. **e**, A representative trace of gel filtration of the CED of Piezo1. The molecular weights are labelled. Protein samples of the indicated fractions were subjected to SDS–PAGE and Coomassie blue staining (bottom). **f**, Transparent surface representation of the segmented density map of the cap, superimposed with the trimeric CED crystal structure. The trimeric CEDs are coloured as in **d**.



Extended Data Figure 7 | Local resolution map of the final density map. **a**, The final 3D density map of Piezo1 is coloured according to the local resolutions estimated by the software of ResMap. The density map is shown in three different views (top, bottom and side, respectively). **b**, The final 3D density map (transparent) is superimposed with a poly-alanine model and the crystal structure of the trimeric CED. Three protomers are coloured cyan, purple and green, respectively.



Extended Data Figure 8 | Density connections between the transmembrane helices and between the helices in the compact CTD. **a**, Alanine models of five representative pairs of transmembrane helices are displayed with their densities (mesh) superimposed. The transmembrane region is highlighted by a light purple shade with the intracellular and extracellular sides indicated. **b**, An alanine model of the anchor motif with its density superimposed (mesh). Four helices ($\alpha1^{anchor}$–$\alpha4^{anchor}$) connecting PH1 and OH are labelled. The transmembrane region is highlighted by a light purple shade with the intracellular and extracellular sides indicated. **c**, An alanine model of the last four helices ($\alpha1^{CTD}$–$\alpha4^{CTD}$) of the trimeric CTD, superimposed with the density of the CTD (mesh).



Extended Data Figure 9 | Secondary structure analyses of the C-terminal segments of Piezo1 proteins from different species. Sequence alignment of the C-terminal regions of Piezo1 from different species. The alignment was performed using Clustal W2 (ref. 71). The anchor motif and the CTD are highlighted in green and pink, respectively. For clarification, the sequences of CEDs were omitted and are indicated by red dashed lines. Secondary structures (α-helices) predicted with PredictProtein[72] are shown as black lines. Transmembrane segments were predicted using multiple web servers including Topcons[73] (green lines), TMHMM2 (ref. 74) (blue lines), HMMTOP[75] (orange lines) and Phobius[76] (pink lines).

Extended Data Table 1 | Statistics of data collection and structure refinement.

	Native	I-SAD
Data collection		
Diffraction beam	SSRF BL17U	BSRF 4W1B
Space Group	P213	P213
Unit Cell (Å)	a=b=c=89.492	a=b=c=89.766
Wavelength (Å)	0.979	1.700
Resolution (Å)	50.00-1.45 (1.50-1.45)	50.00-2.26 (2.34-2.26)
Rmerge (%)	9.2 (48.4)	25.7 (90.2)
I/σ	18.7 (4.0)	21.2 (2.1)
Completeness (%)	99.9 (100.0)	98.3 (82.8)
Redundancy	7.2 (7.1)	33.9 (7.4)
Wilson B factor (Å^2)	13.3	27.20
Refinement		
R_{work} (%)	15.15	
R_{free} (%)	17.59	
No. atoms		
Alll	2159	
Side chains	899	
Main chains	908	
Macromolecules	1807	
Solvent	352	
Average B-factor		
All	18.8	
Side chains	18.2	
Main chains	14.9	
Macromolecule	16.5	
Solvent	30.2	
RMS (bonds)	0.006	
RMS (angles)	1.065	
Ramachandran plot (%)		
Favored	98.21	
Allowed	1.79	
Outliers	0	

Values in parentheses are for the highest resolution shell. $R_{merge} = \Sigma_h\Sigma_i|I_{h,i}-I_h|/\Sigma_h\Sigma_i I_{h,i}$, where I_h is the mean intensity of the i observations of symmetry-related reflections of h. $R = \Sigma||F_{obs}|-|F_{calc}||/\Sigma|F_{obs}|$, where F_{calc} is the calculated protein structure factor from the atomic model (R_{free} was calculated with 5% of the reflections selected). I-SAD, single-wavelength anomalous dispersion of I atoms; BSRF, Beijing Synchrotron Radiation Facility; SSRF, Shanghai Synchrotron Radiation Facility.

LETTER

doi:10.1038/nature15702

Episodic molecular outflow in the very young protostellar cluster Serpens South

Adele L. Plunkett[1], Héctor G. Arce[1], Diego Mardones[2], Pieter van Dokkum[1], Michael M. Dunham[3], Manuel Fernández-López[4], José Gallardo[5] & Stuartt A. Corder[5]

The loss of mass from protostars, in the form of a jet or outflow, is a necessary counterpart to protostellar mass accretion[1,2]. Outflow ejection events probably vary in their velocity and/or in the rate of mass loss. Such 'episodic' ejection events[3] have been observed during the class 0 protostellar phase (the early accretion stage)[4–10], and continue during the subsequent class I phase that marks the first one million years of star formation[11–14]. Previously observed episodic-ejection sources were relatively isolated; however, the most common sites of star formation are clusters[15]. Outflows link protostars with their environment and provide a viable source of the turbulence that is necessary for regulating star formation in clusters[3], but it is not known how an accretion-driven jet or outflow in a clustered environment manifests itself in its earliest stage. This early stage is important in establishing the initial conditions for momentum and energy transfer to the environment as the protostar and cluster evolve. Here we report that an outflow from a young, class 0 protostar, at the hub of the very active and filamentary Serpens South protostellar cluster[16–18], shows unambiguous episodic events. The $^{12}C^{16}O$ ($J = 2-1$) emission from the protostar reveals 22 distinct features of outflow ejecta, the most recent having the highest velocity. The outflow forms bipolar lobes—one of the first detectable signs of star formation—which originate from the peak of 1-mm continuum emission. Emission from the surrounding $C^{18}O$ envelope shows kinematics consistent with rotation and an infall of material onto the protostar. The data suggest that episodic, accretion-driven outflow begins in the earliest phase of protostellar evolution, and that the outflow remains intact in a very clustered environment, probably providing efficient momentum transfer for driving turbulence.

We used the Atacama Large Millimeter/sub-millimeter Array (ALMA) in Chile to observe the $J = 2-1$ emission line of carbon monoxide isotopologues (^{12}CO, ^{13}CO and $C^{18}O$) near the class 0 source CARMA-7 (hereafter C7), in the young protostellar cluster Serpens South. C7 is the strongest of several millimetre-wavelength continuum sources that are densely packed within Serpens South, located at a distance of 415 parsecs (pc) from Earth[19]. Its relative proximity to Earth allows for observations with high spatial resolution; our observations resolve features with physical sizes of greater than about 370 astronomical units (AU).

The ^{12}CO emission extends north–south of C7, spanning about 80″ (or 0.16 pc) along an axis with a position angle of roughly 4° (Fig. 1). The emission is clumpy, and the strongest emission features to the north (south) are mostly redshifted (blueshifted), relative to the systemic cloud velocity (V_c) of 8 km s^{-1} (refs 20, 21). The emission features near the origin are only around 1–2″ (about 400–800 AU) wide, and the width increases to about 8″ (roughly 3,000 AU) at the widest point. The opening angle of the emission decreases with velocity, with a maximum of about 23° (at 10″, or 0.02 pc, from the source) at velocities of a few kilometres per second, and a minimum of about 10° at the same distance and higher velocities. Figure 2a shows the position–velocity diagram, with a saw-like pattern along the extent of the ^{12}CO emission; and emission features corresponding to the highest velocities ($|V_{LSR} - V_c| = \sim 20$ km s^{-1}, where V_{LSR} is the local standard of rest velocity) are found closest to C7 (Fig. 2a, b). The ^{12}CO emission is optically thick—especially, according to our data, near the cloud velocity—and therefore it traces outflow features with velocities greater than a few kilometres per second with respect to V_c.

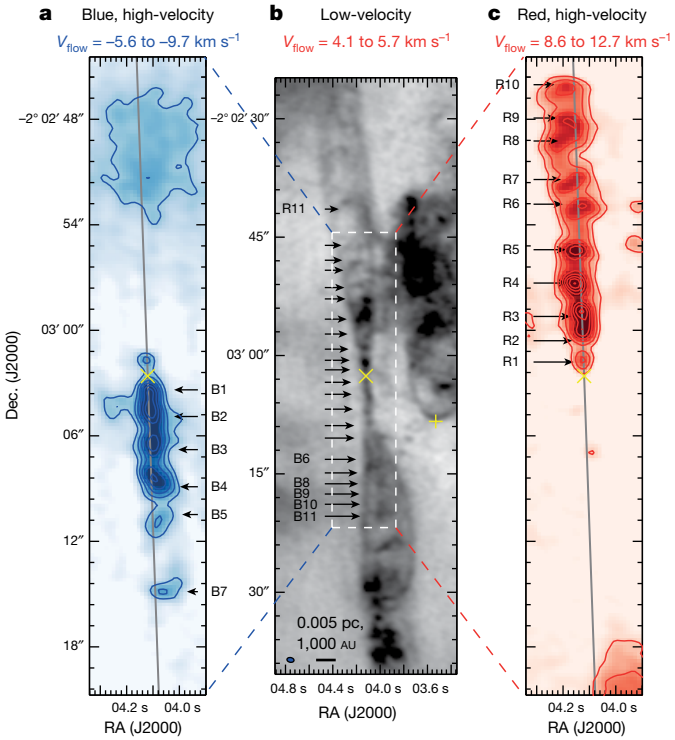


Figure 1 | ^{12}CO molecular outflow emission centered at the class 0 protostar CARMA-7 (C7). C7 is marked by the yellow cross at right ascension RA = 18 h 30 min 04.1 s, declination dec. = −02° 03′ 02.6″. The numbers on the *x* axes are truncated to show seconds only, omitting hours and minutes for brevity. The *y* axes are likewise simplified. **a**, **c**, High-velocity blueshifted and redshifted channels, respectively. **b**, Low-velocity channels, to show the cavity surrounding collimated ejecta. Contours in **a** and **c** begin with 8σ and increment by 4σ and 8σ, respectively. Labels B1–B11 and R1–R11 indicate 22 ejecta features. The grey lines mark the 4° position angle of the C7 outflow lobes. The yellow 'plus' symbol marks a neighbouring protostar, CARMA-6 (ref. 21), which provides contaminating emission, especially for the blueshifted southern outflow lobe.

[1]Astronomy Department, Yale University, New Haven, Connecticut 06511, USA. [2]Departamento de Astronomía, Universidad de Chile, Casilla 36-D, Santiago, Chile. [3]Harvard-Smithsonian Center for Astrophysics, 60 Garden Street, MS 78, Cambridge, Massachusetts 02138, USA. [4]Instituto Argentino de Radioastronomía, CCT-La Plata (CONICET), C.C.5, 1894, Villa Elisa, Argentina. [5]Joint ALMA Observatory, Av. Alonso de Córdova 3107, Vitacura, Santiago, Chile.

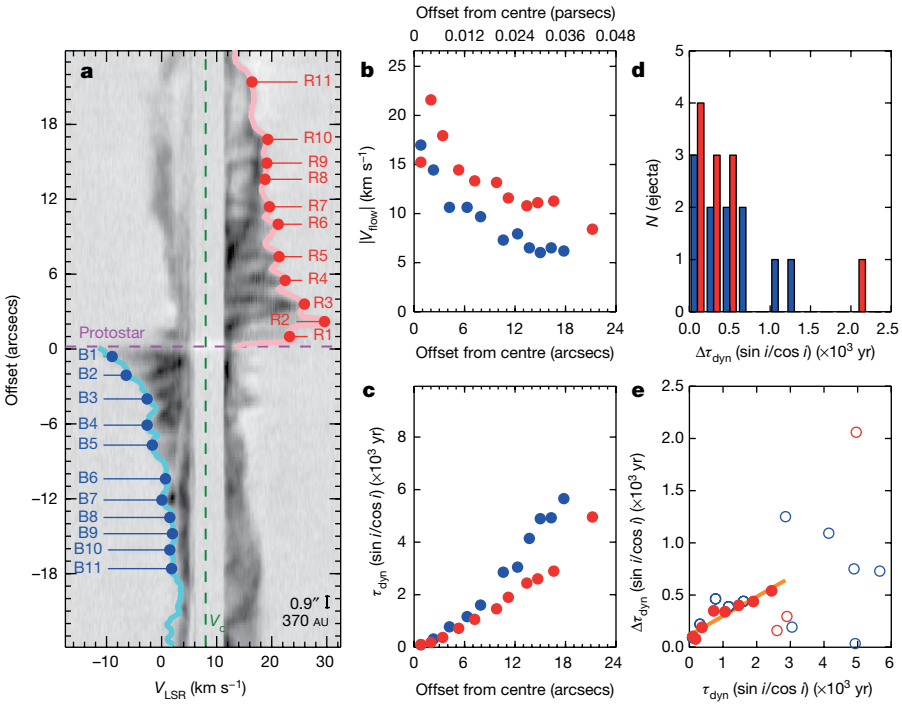


Figure 2 | Outflow ejecta from C7. a, Position–velocity diagram along the outflow axis (see Fig. 1). Points correspond to velocity maxima where we identified outflow knots. Northern emission features are mostly redshifted and are shown in red; southern emission features are mostly blueshifted and are shown in blue. The scale bar shows 370 AU, or 0.9″. The dashed pink line marks the location of the protostar; the dashed green line shows the cloud central velocity, V_c, in the same units as those of V_{LSR}. **b**, Knot velocity (V_{flow}) versus distance relative to C7 (in arcsecs or parsecs). Blue (red) points mark southern (northern) features, as in **a**. **c**, Dynamical timescales (τ_{dyn}) for each knot, with no correction for inclination angle. **d**, Histogram showing the number (N) of ejecta that have been emitted at the given times since the previous ejection ($\Delta\tau_{dyn}$), with 200-year bins. **e**, $\Delta\tau_{dyn}$ as a function of τ_{dyn} for northern (red) and southern (blue) knots. Recent northern ejecta (solid points) are fit with a linear trend (orange line).

The $C^{18}O$ emission is optically thin and therefore probes deeper than does the ^{12}CO emission, to trace denser material that is closer to the protostar (see Fig. 3 and the channel maps in Extended Data Fig. 1). Together, these molecular lines and continuum (Extended Data Fig. 2) trace two related components of the protostellar system: the outflow and the envelope. Material accretes onto a protostar from an infalling envelope via a disk, with the envelope providing the main mass reservoir for the star. While the protostar is still obscured by the surrounding envelope, a bipolar outflow represents one of the first observational signs of star formation, and it carries away mass and angular momentum from the system.

Our observations show two molecular outflow lobes emanating from the C7 envelope, and we conclusively identify an outflow-driving source in this region. When this region was studied with lower-resolution CO observations[21,22], prevalent outflow emission from several young sources appeared to coincide. However, the ^{12}CO emission traces cool (less than about100 K) swept-up outflow material and provides a record of the timing history of mass-loss events for a given source. The C7 outflow comprises cavity walls that surround 22 knots (observed clumps of emission from a single ejection event), 11 to the north and 11 to the south, within 24″ of the source. Beyond this distance, we see outflow morphology that can be attributed to C7, but there is contaminating cloud emission to the north and an interfering outflow to the south (driven by a protostar southwest of C7; Extended Data Fig. 2). The outflow's high collimation, and the presence of redshifted and blueshifted emission coinciding along the

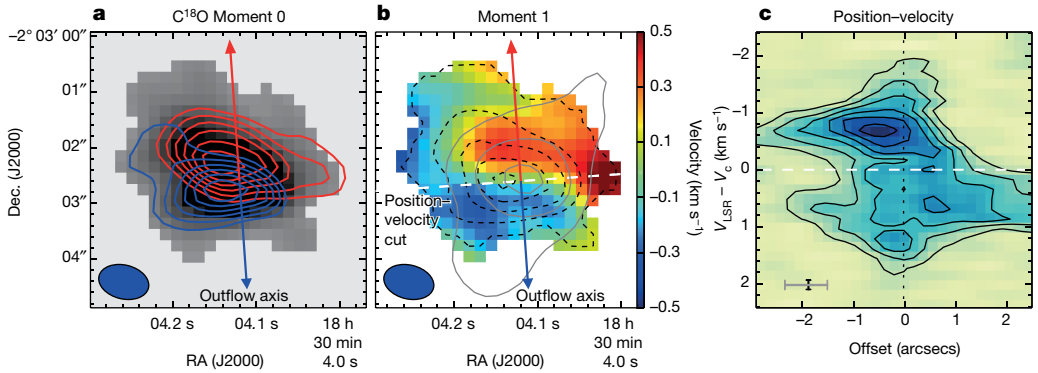


Figure 3 | Protostellar envelope. a, Integrated $C^{18}O$ intensity (moment 0, greyscale) >13σ, with $|V_{LSR} - V_c| < 3$ km s^{-1}. Blue (red) contours represent blueshifted (redshifted) channels, with $0.4 < |V_{LSR} - V_c| < 3$ km s^{-1}, beginning at 30% of peak integrated intensity (165 mJy beam^{-1} km s^{-1} and 155 mJy beam^{-1} km s^{-1}, respectively), with increments of 10% of peak. The blue oval represents the beam size. **b**, Intensity-weighted mean velocity (moment 1, colour scale). Black dashed contours show integrated intensity (greyscale in **a**) with 8σ steps. Grey contours show 15%, 30% and 80% of peak continuum emission (93.9 mJy beam^{-1}). **c**, Position–velocity perpendicular to the outflow axis. Contours show levels of 4σ of the position–velocity intensity. Spatial and velocity resolution elements are shown with grey and black (solid) bars, respectively.

line of sight near the protostar north and south, are consistent with the main outflow axis being oriented nearly in the plane of the sky. Low-velocity redshifted and blueshifted emissions to the south and north are contributed by cavities surrounding the high-velocity jet-like emission, and the jet may precess slightly, given the slight wiggle in the knots shown in Fig. 1a, c.

Clumpy ^{12}CO emission suggests an episodic ejection mechanism, rather than a smooth outflow. Decreasing knot velocities with distance from C7 are consistent with the existence of jet-entrained material that is slowed down by drag because of interaction with the surrounding medium, and/or with the existence of intrinsically variable ejections[23–25]; both probably contribute to the position–velocity trend seen here. The 'superjet' HH34 (driven by the class I source HH34 IRS) also shows a velocity decrease[11], which is proposed to be caused by the drag-induced slowing of jet-entrained material. However, the shapes of the position–velocity curves for HH34 and C7 differ, a difference that may be explained by the relative ages and precession of the sources. The initial C7 ejecta probably cleared some of the dense ambient material, reducing the drag forces for later ejecta following closely behind and in line with previous ejecta. HH34 is more evolved and is precessing to a greater extent, so ejecta seem to be more directly exposed to ambient material, which has not yet been disturbed by previous ejections.

We also find that the velocities of southern (blueshifted) knots from C7 are consistently lower than the velocities of northern (redshifted) knots at comparable distances. This may be evidence for an inhomogeneous ambient cloud medium, such that the southern knots are being slowed down by a denser environment. Alternatively, the jets may be intrinsically variable upon ejection from opposite sides of the disk. It is also possible that the outflow lobes have different inclination angles with respect to the plane of the sky, so that the line-of-sight velocities to the north and south differ. C7 may precess slightly, given that blueshifted emission near C7 shifts to being predominantly red farther from the source.

In Fig. 2c we show dynamic timescales for each of the identified ejecta, ranging from 100 years to 6,000 years (for knots within 24″, or 10,000 AU, of the source). The dynamic timescale for each ejection is given by $\tau_{dyn} = D/V_{flow}$ (cos i/sin i), where D is the distance between an outflow knot and the driving source, V_{flow} is the velocity (along the line of sight) of the knot, and i is the inclination of the outflow with respect to the line of sight. Uncertainties arise because we do not know the inclination angle, and because we assume that the knots travel with constant velocity from the time of their launch. If a jet is launched from the disk[3], then the longest timescale of an (unimpeded) outflow ejection should be a lower limit for the formation time of the disk. The longest timescale of a northern ejection is about 5,000 years; correcting for an inclination angle nearly in the plane of the sky, this could be smaller by a factor of about 10 (for $i = \sim 85°$) or more, which is consistent with the youthfulness of the source. Given that southern knots appear to have lower velocities than northern knots, the timescales of the southern (blueshifted) knots are longer on average than the northern (redshifted) knots.

We quantify the episodic nature of the ejections, and corresponding accretion and/or disk instabilities[3,26], on the basis of the difference in timescales, $\Delta\tau_{dyn}$, for successive ejection events (Fig. 2d). Because of the contamination from the surrounding outflow emission to the south, we base the following calculations on the northern lobe only (within 24″ of C7). In Fig. 2e, we see that seven knots to the north show linearly increasing $\Delta\tau_{dyn}$ as a function of τ_{dyn}, with $\Delta\tau_{dyn}$ ranging from 80 years to 540 years, and a mean $\Delta\tau_{dyn}$ of 310 ± 150 years. These seven knots are the most recently ejected to the north, with a τ_{dyn} of less than 2,400 years (uncorrected for inclination angle). Several modes of velocity variability have been suggested for protostellar jets[13,27], with periods of a few tens, a few hundreds, and a few thousands of years; in the case of a class 0 source, and assuming that C7 has an inclination approximately in the plane of the sky, we are probably witnessing ejecta that are associated mostly with the shorter period modes.

We estimate that, within some 3,000 years, the farthest (slowest) ejecta in the north and south will have been overcome by each of the following (faster) ejecta, if ejecta travel with constant velocities (an admittedly simple assumption). These interactions will produce bright shocks along the outflow. 'Snapshots' of shocks in outflows can be seen in the emission of molecular hydrogen (H_2)[28], which has a higher excitation temperature than does ^{12}CO but cools quickly. Two H_2 bow-shaped shock structures, corresponding to faint, low-velocity ^{12}CO emission lines 38″ (0.08 pc) and 47″ (0.09 pc) north and south of C7, respectively, are seen in the Spitzer 4.5-μm map of the region. These structures may be evidence of the first occurrence of a longer-period mode (of a few hundred years or more), where faster ejecta recently overcame slower ejecta. We propose that frequent ejection bursts during the class 0 phase entrain molecular outflow material, which therefore appears clumpy, creating observable shocks when the ejecta overtake previous ejecta.

Alternatively, if the position–velocity trend provides evidence for an interaction between ejecta and the environment, then the drag-induced momentum loss along the outflow signifies momentum transfer to the environment—an important mechanism that is proposed to drive turbulent motions in a clustered region[3]. We are carrying out further analysis of momentum injection along the span of the outflow at such an early stage, taking into account velocity-dependent opacity of the ^{12}CO line and varying excitation temperatures throughout the outflow.

The C^{18}O envelope seems to be oriented perpendicular to the outflow axis, with its major axis approximately east–west. Elongated blueshifted and redshifted C^{18}O emissions east and west of C7, respectively, are evidence of a non-spherical, rotating envelope. Blueshifted and redshifted peaks of high-velocity emission near C7 to the south and north, respectively, are consistent with infall motion onto a disk that is slightly inclined[29].

Two features in the C^{18}O position–velocity diagram are representative of some contribution from unresolved Keplerian rotation (on scales of less than ~400 AU): larger velocities at smaller distances, and position–velocity intensity peaks offset bluewards and redwards from the line $V_{LSR} = V_c$. The position–velocity structure for C7 is consistent with a combination of rotation and infall on a slightly inclined disk, as shown in models[29] and sketched in Extended Data Fig. 3. However, the C^{18}O position–velocity diagram (Fig. 3c) also shows some deviations from models of a rotating, infalling envelope: first, the blueshifted peak is stronger than the redshifted peak; second, redshifted emission with velocities $V_{LSR} - V_c = \sim 0.5 - 1$ km s^{-1} coincides with strong blueshifted emission at an offset of about −2″; and third, redshifted extended emission west of C7 probably contaminates the C7 envelope emission. The outflow and envelope that we observe here clearly pertain to the same protostar, and higher-resolution observations of the disk and envelope will reveal the jet-launching region and disk-formation mechanisms in this young system.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 16 June; accepted 3 September 2015.

1. Norman, C. & Silk, J. Clumpy molecular clouds—a dynamic model self-consistently regulated by T Tauri star formation. *Astrophys. J.* **238,** 158–174 (1980).
2. Shu, F. H., Adams, F. C. & Lizano, S. Star formation in molecular clouds—observation and theory. *Annu. Rev. Astron. Astrophys.* **25,** 23–81 (1987).
3. Frank, A. *et al.* in *Protostars and Planets VI* (eds Beuther, H. *et al.*) 451–474 (Univ. Arizona, 2014).
4. Gueth, F. & Guilloteau, S. The jet-driven molecular outflow of HH 211. *Astrophys. J.* **343,** 571–584 (1999).
5. Lee, C.-F., Mundy, L. G., Reipurth, B., Ostriker, E. C. & Stone, J. M. CO outflows from young stars: confronting the jet and wind models. *Astrophys. J.* **542,** 925–945 (2000).
6. Lee, C.-F. *et al.* HH 212: submillimeter array observations of a remarkable protostellar jet. *Astrophys. J.* **659,** 499–511 (2007).
7. Lee, C.-F. *et al.* Submillimeter arcsecond-resolution mapping of the highly collimated protostellar jet HH 211. *Astrophys. J.* **670,** 1188–1197 (2007).

8. Santiago-García, J., Tafalla, M., Johnstone, D. & Bachiller, R. Shells, jets, and internal working surfaces in the molecular outflow from IRAS 04166+2706. *Astrophys. J.* **495,** 169–181 (2009).
9. Hirano, N. *et al.* Extreme active molecular jets in L1448C. *Astrophys. J.* **717,** 58–73 (2010).
10. Loinard, L. *et al.* ALMA and VLA observations of the outflows in IRAS 16293–2422. *Mon. Not. R. Astron. Soc.* **430,** L10–L14 (2013).
11. Cabrit, S. & Raga, A. Theoretical interpretation of the apparent deceleration in the HH 34 superjet. *Astrophys. J.* **354,** 667–673 (2000).
12. Goodman, A. A. & Arce, H. G. PV Cephei: young star caught speeding? *Astrophys. J.* **608,** 831–845 (2004).
13. Ioannidis, G. & Froebrich, D. YSO jets in the galactic plane from UWISH2—II. Outflow luminosity and length distributions in Serpens and Aquila. *Mon. Not. R. Astron. Soc.* **425,** 1380–1393 (2012).
14. Arce, H. G. *et al.* ALMA observations of the HH 46/47 molecular outflow. *Astrophys. J.* **774,** 39 (2013).
15. Lada, C.-J. & Lada, E. A. Embedded clusters in molecular clouds. *Mon. Not. R. Astron. Soc.* **41,** 57–115 (2003).
16. Gutermuth, R. A. *et al.* The Spitzer Gould belt survey of large nearby interstellar clouds: discovery of a dense embedded cluster in the Serpens-Aquila Rift. *Astrophys. J.* **673,** L151–L154 (2008).
17. Tanaka, T. *et al.* The dynamical state of the Serpens South filamentary infrared dark cloud. *Astrophys. J.* **778,** 34 (2013).
18. Nakamura, F. *et al.* Cluster formation triggered by filament collisions in Serpens South. *Astrophys. J.* **791,** L23 (2014).
19. Dzib, S. *et al.* VLBA determination of the distance to nearby star-forming regions. IV. A preliminary distance to the proto-Herbig AeBe star EC 95 in the Serpens core. *Astrophys. J.* **718,** 610–619 (2010).
20. Kirk, H. *et al.* Filamentary accretion flows in the embedded Serpens South protocluster. *Astrophys. J.* **766,** 115–128 (2013).
21. Plunkett, A. L. *et al.* Assessing molecular outflows and turbulence in the protostellar cluster Serpens South. *Astrophys. J.* **803,** 22 (2015).
22. Nakamura, F. *et al.* Molecular outflows from the protocluster Serpens South. *Astrophys. J.* **737,** 56 (2011).
23. Raga, A. C., Binette, L., Canto, J. & Calvet, N. Stellar jets with intrinsically variable sources. *Astrophys. J.* **364,** 601–610 (1990).
24. Suttner, G., Smith, M. D., Yorke, H. W. & Zinnecker, H. Multi-dimensional numerical simulations of molecular jets. *Astron. Astrophys.* **318,** 595–607 (1997).
25. Smith, M. D., Suttner, G. & Yorke, H. W. Numerical hydrodynamic simulations of jet-driven bipolar outflows. *Astron. Astrophys.* **323,** 223–230 (1997).
26. Audard, M. *et al.* in *Protostars and Planets VI* (eds Beuther, H. *et al.*) 387–410 (Univ. Arizona, 2014).
27. Raga, A. C., Velázquez, P. F., Cantó, J. & Masciadri, E. The time-dependent ejection velocity histories of HH 34 and HH 111. *Astrophys. J.* **395,** 647–656 (2002).
28. Teixeira, G. D. C., Kumar, M. S. N., Bachiller, R. & Grave, J. M. C. Molecular hydrogen jets and outflows in the Serpens South filamentary cloud. *Astrophys. J.* **543,** A51 (2012).
29. Oya, Y. *et al.* A substellar-mass protostar and its outflow of IRAS 15398–3359 revealed by subarcsecond-resolution observations of H_2CO and CCH. *Astrophys. J.* **795,** 152 (2014).

Acknowledgements A.L.P. is supported by a National Science Foundation (NSF) Graduate Research Fellowship under grant DGE-1122492; this research was made possible by the US Student Program of Fulbright Chile. H.G.A. receives funding from the NSF under grant AST-0845619. D.M. acknowledges support from CONICYT project PFB-06. M.M.D. acknowledges support from the Submillimeter Array through a postdoctoral fellowship. ALMA is a partnership of the European Space Organization (ESO, representing its member states), NSF (USA) and National Institutes of Natural Sciences (Japan), together with the National Research Council (Canada) and National Security Council and Academia Sinica Institute of Astronomy and Astrophysics (Taiwan), in cooperation with the Republic of Chile. The Joint ALMA Observatory is operated by ESO, Associated Universities Inc. (AUI)/National Radio Astronomy Observatory (NRAO) and National Astronomical Observatory of Japan. The NRAO is a facility of the NSF, operated under cooperative agreement by AUI. This paper makes use of the following ALMA data: ADS/JAO.ALMA 2012.1.00769.S.

Author Contributions A.L.P. led the proposal, observations, analysis and interpretation, and wrote the manuscript. H.G.A. contributed to the analysis and interpretation, and to the manuscript. A.L.P., H.G.A., D.M., M.M.D., J.G. and S.A.C. planned the early stages of the project. D.M., M.M.D., M.F.-L. and J.G. contributed to the analysis and interpretation and commented on the manuscript. P.v.D. contributed to the interpretation and to the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to A.L.P. (adele.plunkett@yale.edu) or H.G.A. (hector.arce@yale.edu).

METHODS

Observations and data analysis. The analysis is based on ALMA Cycle 1 observations made with the 12-metre and 7-metre arrays during March 2014 and January to June 2014. The observed mosaics span 2′ × 3′ and consist of 137 and 53 pointings, separated by 15″ and 26″, made by the 12-metre and 7-metre arrays, respectively. Here, we focus on the roughly 90″ × 20″ region centred at RA = 18 h 30 min 04.1 s, dec. = −02° 03′ 02.6″.

The ALMA correlator was configured in the frequency division mode (FDM) of band 6 with four independent spectral windows: one window was assigned to the $J = 2-1$ energy-level transition of each of the spectral lines ^{12}CO (230.538 GHz), ^{13}CO (220.399 GHz) and C^{18}O (219.560 GHz), and the fourth was dedicated to continuum at 231.450 GHz. The bandwidth for each spectral-line window was 234.375 MHz, and the continuum window had a bandwidth of 468.750 MHz. To make a continuum-emission map, we included line-free channels in all spectral windows, resulting in a total continuum bandwidth of 996 MHz. The molecular line data for ^{12}CO and C^{18}O, as well as the continuum, are included in the present analysis.

We performed calibration of the raw visibility data with the common astronomy software application (CASA, version 4.3.0), using the standard reduction script for Cycle 1. We assigned weights to the measurement sets using the task 'statwt' and combined the calibrated 12-metre and 7-metre array UV data using the task 'concat'.

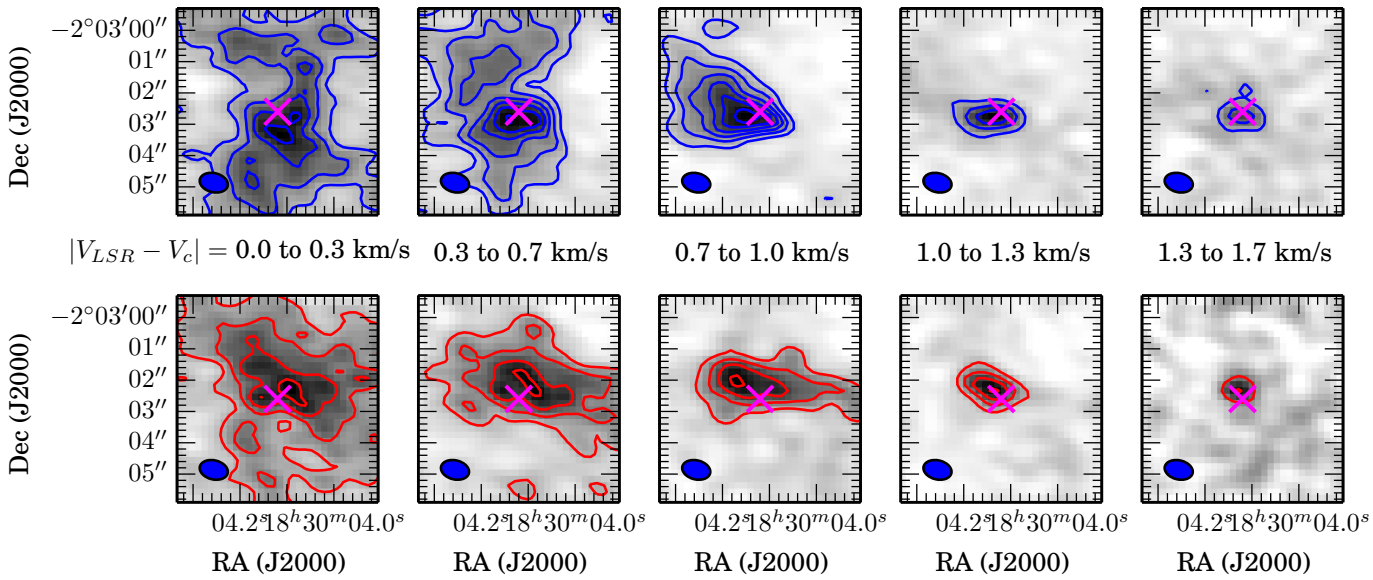
We created image cubes for each molecular line, as well as the continuum image, by first applying a Fourier transform to the calibrated data, producing an intermediate ('dirty') image. Using the intermediate image, we drew masks around the emission features, and these masks were used in an interactive 'clean' process to deconvolve the telescope point-spread function. We used Briggs weighting with a robust parameter of 0.5, and we imaged with a cell size of 0.3″ and a spectral (velocity) resolution of 0.16 km s^{-1}. Finally, we subtracted continuum emission from the spectral-line data using the task 'imcontsub'.

The resulting beam sizes for the ^{12}CO and C^{18}O data cubes are 0.9″ × 0.6″ (with position angles of 79.7° and 76.3° for ^{12}CO and C^{18}O, respectively). The root-mean-squared (r.m.s.) noise levels are 9 mJy beam^{-1} channel^{-1} and 8 mJy beam^{-1} channel^{-1} respectively, with channel widths of 0.16 km s^{-1}. The r.m.s. noise level for the continuum is 0.2 mJy beam^{-1} near the edge of the region presented here, with an upper-limit r.m.s. noise level of 0.3 mJy beam^{-1} within 30″ of the strong continuum emission.

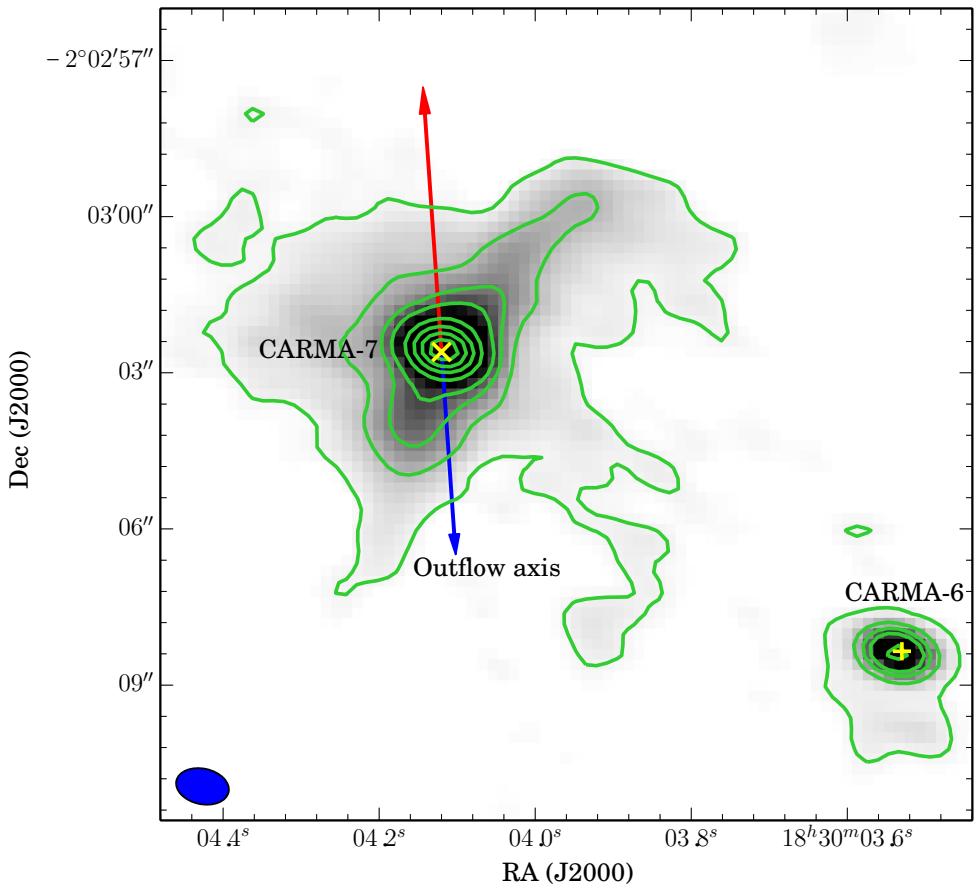
C^{18}O channel maps. The C^{18}O emission, shown in Extended Data Fig. 1, is concentrated where the northern redshifted and southern blueshifted ^{12}CO emissions meet. The C^{18}O morphology changes from extended at the lowest velocities ($|V_{LSR} - V_c| = 0-0.7$ km s^{-1}) to compact and oriented approximately north–south, or coincident with the outflow axis, at higher velocities ($|V_{LSR} - V_c| = 1.3-1.7$ km s^{-1}). At intermediate velocities ($|V_{LSR} - V_c| = 0.6-1$ km s^{-1}), elongated blueshifted and redshifted emissions are seen east and west of C7, respectively. A shell in the C^{18}O emission in the south, and less noticeably in the north, is seen bisected by the ^{12}CO axis in Fig. 3a, b. This is similar to the situation with the protostar HH212 (ref. 30), and in both cases material originally in the envelope is probably swept up to form the cavity. It may be too early for the outflow to have a noticeable impact on the infall and rotation motions of the envelope.

Continuum emission. The continuum emission peaks in our map at RA = 18 h 30 min 04.1 s, dec. = −02° 03′ 02.6″ (see Extended Data Fig. 2), with an intensity of 93.9 mJy beam^{-1}, and this coincides with the centre of the C^{18}0 emission (Fig. 3). Although the highest-intensity continuum emission (greater than ~50σ) is concentrated and can be fit well with a two-dimensional Gaussian curve, the weaker (yet statistically significant) continuum emission is elongated northwest–southeast. Additional continuum emission from the nearby protostar CARMA-6 may contribute to the extended continuum emission, and molecular outflow emission is also associated with this source (although not shown here).

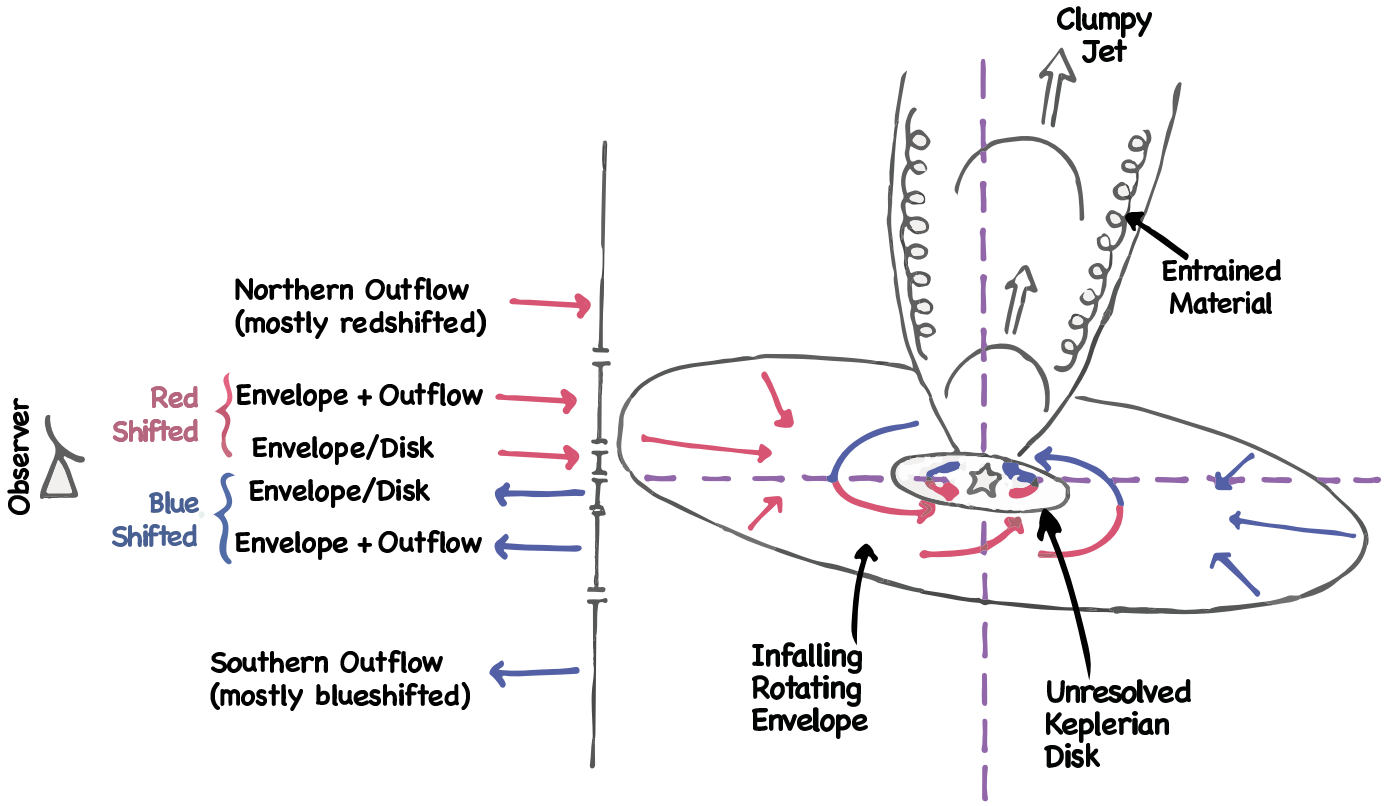
30. Lee, C.-F. *et al.* ALMA results of the pseudodisk, rotating disk, and jet in the continuum and HCO$^+$ in the protostellar system HH 212. *Astrophys. J.* **786,** 114 (2014).



Extended Data Figure 1 | $C^{18}O$ emission from the protostellar source C7. Top row, blueshifted emission; bottom row, redshifted emission; velocity increases from left to right. Contours begin at 4σ and increment by 4σ. Specific velocity ranges ($|V_{LSR} - V_c|$, or velocity relative to cloud velocity) are given for each column. Each panel shows integrated emission from two channels. The location of peak continuum emission is marked with a magenta cross.



Extended Data Figure 2 | 1-mm continuum emission near the sources CARMA-7 (RA = 18 h 30 min 04.1 s, dec. = −02° 03′ 02.6″) and CARMA-6 (RA = 18 h 30 min 03.5 s, dec. = −02° 03′ 08.4″). Contours show 10σ, 30σ, 50σ and 70σ, followed by increments of 50σ. Near these strong sources, we find the r.m.s. noise to be 0.3 mJy beam^{-1}.



Extended Data Figure 3 | Cartoon depiction of a protostellar system, showing the outflow (^{12}CO emission), envelope (C^{18}O emission) and disk (unresolved). Contributions to blueshifted and redshifted molecular line emission are indicated along the outflow and envelope, assuming that the outflow is nearly in the plane of the sky with respect to the observer.

LETTER

doi:10.1038/nature15735

Hong–Ou–Mandel interference of two phonons in trapped ions

Kenji Toyoda[1], Ryoto Hiji[1], Atsushi Noguchi[1]† & Shinji Urabe[1]

The quantum statistics of bosons and fermions manifest themselves in the manner in which two indistinguishable particles interfere quantum mechanically. When two photons, which are bosonic particles, enter a beam-splitter with one photon in each input port, they bunch together at either of the two output ports. The corresponding disappearance of the coincidence count is the Hong–Ou–Mandel effect[1]. Here we show the phonon counterpart of this effect in a system of trapped-ion phonons, which are collective excitations derived by quantizing vibrational motions that obey Bose–Einstein statistics. We realize a beam-splitter transformation of the phonons by employing the mutual Coulomb repulsion between ions, and perform a two-phonon quantum interference experiment using that transformation. We observe an almost perfect disappearance of the phonon coincidence between two ion sites, confirming that phonons can be considered indistinguishable bosonic particles. The two-particle interference demonstrated here is purely a quantum effect, without a classical counterpart, hence it should be possible to demonstrate the existence of entanglement on this basis. We attempt to generate an entangled state of phonons at the centre of the Hong–Ou–Mandel dip in the coincidence temporal profile, under the assumption that the entangled phonon state is successfully generated if the fidelity of the analysis pulses is taken into account adequately. Two-phonon interference, as demonstrated here, proves the bosonic nature of phonons in a trapped-ion system. It opens the way to establishing phonon modes as carriers of quantum information in their own right[2–4], and could have implications for the quantum simulation of bosonic particles[5,6] and analogue quantum computation via boson sampling[7].

When two photons with the same wave-packet temporal profiles and polarization are made to interfere at a 50:50 beam-splitter, the coincidence between the photon detection at the two output ports disappears. This Hong–Ou–Mandel (HOM) effect[1,8–12] has also been observed for atoms[13,14], and a related effect has been noted in the case of fermions[15,16]. The HOM effect and the underlying mechanisms for the generation and interference of indistinguishable particles not only reveal the fundamental natures of these particles, but also enable large-scale quantum information processing (QIP)[17]. The HOM effect has also been used to generate entanglement between two atomic ions separated by one metre[18].

In research into QIP using trapped ions, phonons have usually played a supporting role in mediating interactions between internal-state qubits or pseudo-spins[19,20]. They are also expected to play a central role in simulating bosonic-particle systems[5,7]. As a crucial step towards phonon-based applications, the coupling of multiple vibrational modes of ions at a single quantum level has been realized[2–4]. Phonon indistinguishability is another key component necessary for these applications, but this has not been explicitly demonstrated previously.

A system of trapped ions presents an ideal environment for QIP and quantum simulation. In this study, the almost perfect matching of radial frequencies among different sites in a linear trap and the near perfect preparation of initial states by sideband cooling assure the exact indistinguishability of the phonons in this system.

For two ions in a linear Paul trap, the hopping Hamiltonian for the local phonon operators is expressed as (see Methods for details)

$$\hat{H}_{\mathrm{I}} = \frac{\hbar\kappa}{2}(\hat{a}_1\hat{a}_2^{\dagger} + \hat{a}_1^{\dagger}\hat{a}_2) \tag{1}$$

Here, κ is the hopping rate of the radial phonons between two sites, $\hbar$ is $h/2\pi$ where h is the Planck constant, and $\hat{a}_i$ and $\hat{a}_i^{\dagger}$ are the annihilation and creation operators of the radial phonons at the ith site, respectively. From this Hamiltonian, the propagator can be calculated and the local phonon operators for sites 1 and 2 in the Heisenberg picture are expressed as (for a detailed derivation, see Methods)

$$\hat{a}_1(t) = \hat{a}_1\cos\frac{\kappa t}{2} - i\hat{a}_2\sin\frac{\kappa t}{2} \tag{2}$$

$$\hat{a}_2(t) = -i\hat{a}_1\sin\frac{\kappa t}{2} + \hat{a}_2\cos\frac{\kappa t}{2} \tag{3}$$

When time $t = T_{\mathrm{hop}}/4$, where $T_{\mathrm{hop}} = 2\pi/\kappa$ is the hopping period, this transformation corresponds to a 50:50 beam-splitter in linear optics.

With the transformation given above, the initial product state $|1\rangle_1|1\rangle_2$ is transformed to $-i(|2\rangle_1|0\rangle_2 + |0\rangle_1|2\rangle_2)/\sqrt{2}$, where $|n\rangle_i$ represents the phonon Fock state of the ith ion with the quantum number n. This (ideal) state is an entangled state (a NOON state with $N = 2$), which is a superposition of states where phonons are bunched at either of the two sites. Thus, the coincidence of phonons between the two sites disappears. See Fig. 1 for a conceptual diagram of the phonon dynamics.

To observe the two-phonon interference, we use the radial modes of two $^{40}Ca^{+}$ ions in a linear Paul trap with secular frequencies $(\omega_x, \omega_y, \omega_z)/2\pi = (3.45, 3.20, 0.11)$ MHz, where x and y are the two radial directions and z is the axial direction. The phonon modes in one radial direction, y, are used here as the modes that will be manipulated and observed in the experiment. The distance between the two ions is 24 μm and the hopping rate $\kappa/2\pi \approx 2$ kHz. Doppler cooling is performed by illuminating the system with a 397-nm laser resonant to the $S_{1/2} \leftrightarrow P_{1/2}$ electronic state transition and an 866-nm laser resonant to $D_{3/2} \leftrightarrow P_{1/2}$. The states $S_{1/2}$ ($m_J = -1/2$) and $D_{5/2}$ ($m_J = -5/2$) are used as the internal ground and excited states for the present experiment, where m_J is the projection of the total electronic angular momentum. Sideband cooling of the motional states and manipulation of the carrier and sideband transitions in the y direction are performed by illuminating the system with a 729-nm laser resonant to the $S_{1/2} \leftrightarrow D_{5/2}$ transition. An 854-nm laser resonant to $D_{5/2} \leftrightarrow P_{3/2}$ is also used as a quenching laser in the sideband cooling and to clear out the population in the excited state. Observation of the internal states is performed using a photomultiplier, by illuminating the ions with the 397-nm and 866-nm lasers. Further details of the experimental procedures are given in Methods and in our previous publications[4,6].

[1]Graduate School of Engineering Science, Osaka University, 1-3 Machikaneyama, Toyonaka, Osaka 560-8531, Japan. †Present address: Research Center for Advanced Science and Technology, The University of Tokyo, 4-6-1 Komaba, Meguro-ku, Tokyo 153-8904, Japan.

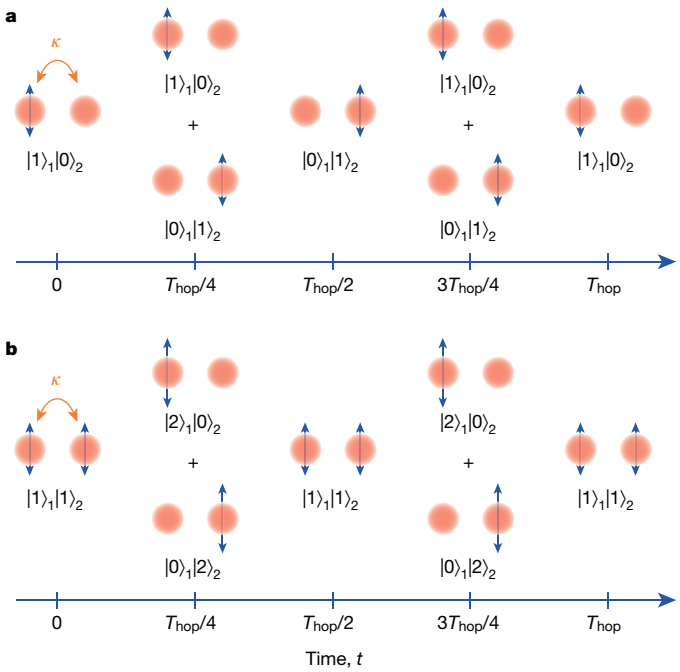


Figure 1 | Conceptual diagrams of phonon hopping dynamics and two-phonon interference. a, Hopping dynamics. An initial Fock state $|1\rangle_1|0\rangle_2$ with one phonon is prepared at $t=0$, and the phonon hops back and forth between the two ion sites with period $T_{\mathrm{hop}}\equiv 2\pi/\kappa$. **b**, Two-phonon interference. An initial Fock state, $|1\rangle_1|1\rangle_2$, with two phonons is prepared. The two phonons are made to hop between two sites, and at $t=T_{\mathrm{hop}}/4$ and $3T_{\mathrm{hop}}/4$, the HOM effect is observed; thus, the phonon population coincidence between the two sites disappears and an entangled state, $(|2\rangle_1|0\rangle_2+|0\rangle_1|2\rangle_2)/\sqrt{2}$, is generated.

We first examined the hopping dynamics[2–4] due to the Hamiltonian $\hat{H}_{\mathrm{I}}$ for an initial Fock state $|g,1\rangle_1|g,0\rangle_2$. Here, $|g(e),n\rangle_i$ represents the basis for the internal ground (excited) state with n local phonons in the ion site i. The experimental procedure is as follows: (1) all the radial vibrational modes of the two ions are cooled to near the ground state via sideband cooling; thus, the initial state $(|g,0\rangle_1|g,0\rangle_2)$ is prepared; (2) ion 1 is irradiated with a π pulse (duration ~19 μs), which is resonant with the blue-sideband transition; hence, the state is transferred to $|e,1\rangle_1|g,0\rangle_2$; (3) immediately after this, the ions are irradiated with an 854-nm (quenching) pulse with 30-μs duration to re-initialize the internal state of ion 1 to the ground state. The expected state after this operation is $|g,1\rangle_1|g,0\rangle_2$, which is used as the initial state for the hopping experiment; (4) a pause with no laser irradiation is permitted to allow the phonon system to undergo hopping; (5) a red-sideband π pulse is applied to the ions to map the manifold $\{|g,0\rangle_i,|g,1\rangle_i\}$ to $\{|g,0\rangle_i,|e,0\rangle_i\}$. The 397-nm laser illumination is then used to record the fluorescence and, hence, the phonon state is estimated. Steps (1)–(5) are repeated with different pause durations (4) to deduce T_{hop}.

Figure 2a shows the result of phonon hopping for two ions with a pause duration of up to 11 ms (circles). The horizontal and vertical axes represent the pause duration and the mean phonon number of ion 1, respectively. The fit to a sinusoidal function with an exponentially decaying envelope (solid curve) gives $\kappa=2\pi\times 2.05\,\mathrm{kHz}$ ($T_{\mathrm{hop}}=489\,\mu\mathrm{s}$) and an e^{-1} decay time of 13.0 ms.

Figure 2b shows the same result (circles) with a magnified horizontal axis scale. According to the fit (solid curve), at $t\approx 57\,\mu\mathrm{s}$ (marked with a red vertical line) the effect of the hopping Hamiltonian corresponds to the transformation produced by a 50:50 beam-splitter. The dashed curve is a simulated result (see Methods) and basically reproduces the experimental finding both qualitatively and quantitatively. In the result shown in Fig. 2b, the maximum value of the phonon population does not reach 1. In addition, the results do not begin from this value at the origin of the horizontal axis. These imperfections are explained in Methods.

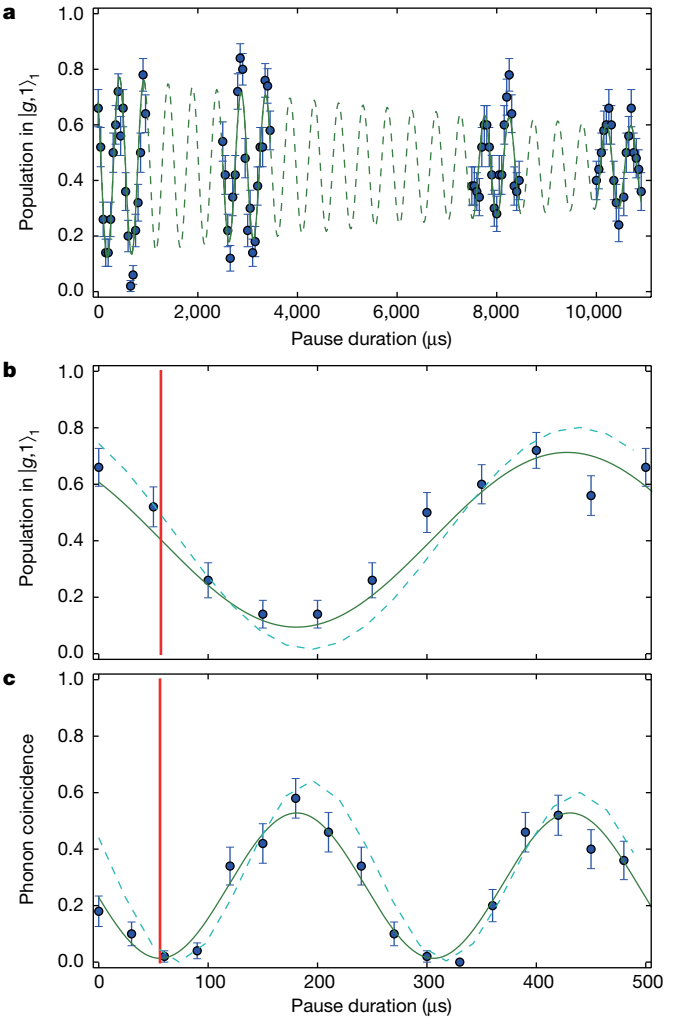


Figure 2 | Experimental results for phonon dynamics. a, Observed phonon hopping dynamics. The horizontal and vertical axes represent the pause duration and phonon population in ion 1, respectively. The circles indicate experimental values, and the combination of solid and dashed curves is a fit with a sinusoidal function with an exponentially decaying envelope. The vertical line at $t=57\,\mu\mathrm{s}$ represents the point for the 50:50 beam-splitter transformation, which is estimated from the fitting. **b**, The same result as previously (**a**) with a magnified horizontal scale, for comparison with the next result (**c**), which shares the same horizontal axis. The circles are experimental values, the solid curve is a sinusoidal fit, and the dashed curve is a simulation result. The vertical line at $t=57\,\mu\mathrm{s}$ represents the point for the 50:50 beam-splitter transformation, which is estimated from the fitting. **c**, Observed coincidence. The coincidence of the internal states after the red-sideband π pulse for mapping is interpreted as the phonon coincidence. The horizontal and vertical axes represent the pause duration and coincidence, respectively. The circles are experimental values, the solid curve is a sinusoidal fit, and the dashed curve is a simulation result. The vertical line at $t\approx 56\,\mu\mathrm{s}$ represents the point for the 50:50 beam-splitter transformation, which is estimated from the fitting. The error bars denote the standard deviation calculated from the variances and covariances of multinomial distributions. The number of measurements per data point is 50.

Next, we observed the coincidence of two phonons and attempted to generate an entangled state, having an ideal form of $(|2\rangle_1|0\rangle_2+|0\rangle_1|2\rangle_2)/\sqrt{2}$. The majority of the experimental procedure is identical to that of the hopping experiment. The differences are that, in step (2), both ions 1 and 2 are irradiated with a blue-sideband π pulse; thus, the initial state $|g,1\rangle_1|g,1\rangle_2$ is prepared after step (3). Further, in step (5), among the three possibilities for a phonon Fock state having two phonons ($\{|g,0\rangle_1|g,2\rangle_2,|g,1\rangle_1|g,1\rangle_2,|g,2\rangle_1|g,0\rangle_2\}$), only $|g,1\rangle_1|g,1\rangle_2$ is transferred by a red-sideband π pulse to a state having two internal excitations (that is, $|e,0\rangle_1|e,0\rangle_2$). Then, the 397-nm

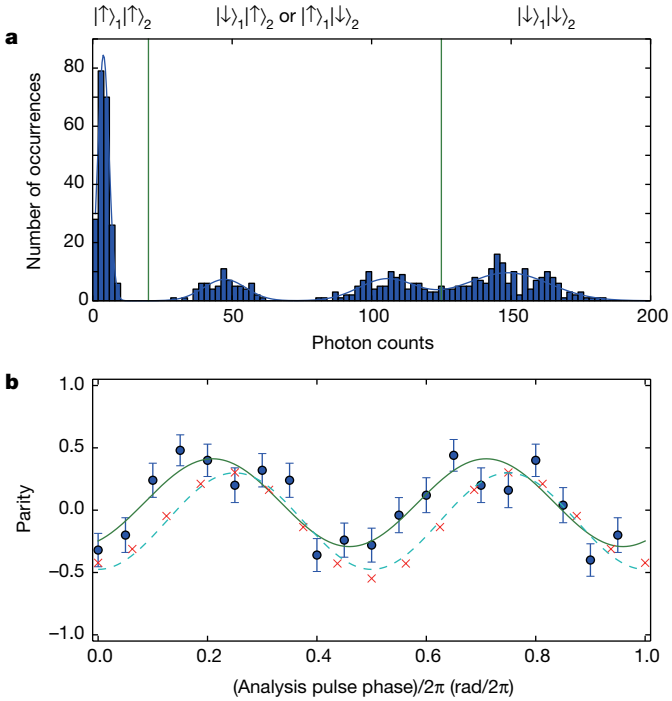


Figure 3 | Measurement of fidelity of the $(|2\rangle_1|0\rangle_2 + |0\rangle_1|2\rangle_2)/\sqrt{2}$ state. **a**, Observed ion fluorescence. The horizontal and vertical axes represent the fluorescence photon counts and the number of occurrences of each photon count. The vertical lines represent the threshold levels used for discrimination of the photon counts. The solid curve is a fit with the sum of four Gaussians. **b**, Measurement of parity against the phase of the analysis $\pi/2$ pulse. The circles indicate experimental values and the solid curve is a sinusoidal fit. The crosses are simulation results, and the dashed curve is a sinusoidal fit to the simulated data. The simulation result reproduces the experimental result well. The error bars denote the standard deviation calculated from the variances and covariances of multinomial distributions. The number of measurements per data point is 50.

laser illuminates the system in the same way as above, and the coincidence of the internal excitations, that is, the probability that both of the ions are shelved to $D_{5/2}$, is estimated from the fluorescence. This coincidence of internal excitations is finally interpreted as the phonon coincidence between the two sites before the application of the red-sideband π pulse.

Figure 2c shows the results of the phonon coincidence between the two sites (circles). According to the fit with a sinusoidal function (solid curve), at $t \approx 56\,\mu s$ (marked with a red vertical line) the effect of the hopping Hamiltonian amounts to the transformation produced by a 50:50 beam-splitter. This point in time, $t \approx 56\,\mu s$, is very close to the corresponding time point in Fig. 2b ($t \approx 57\,\mu s$), hence, it is safe to say that the two results are consistent in this regard. We can see a dip at this point, which indicates that both of the ion sites are not simultaneously occupied by phonons. The almost perfect disappearance of the coincidence guarantees almost perfect interference between the two phonons. The dashed curve is obtained through simulation (see Methods) and also reproduces the experimental result well.

The phonon state at the dip is expected to be an entangled state $(|2\rangle_1|0\rangle_2 + |0\rangle_1|2\rangle_2)/\sqrt{2}$, although we cannot ignore the populations in states such as $|0\rangle_1|0\rangle_2$, $|0\rangle_1|1\rangle_2$ and $|1\rangle_1|0\rangle_2$, which do not contribute to the coincidence. These states may be mixed because of imperfect preparation of the initial state, $|1\rangle_1|1\rangle_2$, by blue-sideband pulses (see Methods for details of this imperfection). In order to confirm the generation of the entangled state, we estimated its fidelity using optical pulses to transform the state to $(|\uparrow\rangle_1|\uparrow\rangle_2 + |\downarrow\rangle_1|\downarrow\rangle_2)/\sqrt{2}$, where $\{|\downarrow\rangle_i, |\uparrow\rangle_i\} \equiv \{|g,0\rangle_i, |e,1\rangle_i\}$. The density-matrix components after this transformation were obtained, which were used to calculate the fidelity (see Methods for details).

Figure 3a shows the measured populations in the internal states for estimation of the diagonal components of the density matrix, $\rho_{\uparrow\uparrow\uparrow\uparrow}$ and $\rho_{\downarrow\downarrow\downarrow\downarrow}$. If the ions are in $|\uparrow\rangle_1|\uparrow\rangle_2$, no fluorescence should be recorded, and if the ions are in $|\downarrow\rangle_1|\downarrow\rangle_2$, fluorescence from both of the ions should be recorded. Therefore, the sum of the peaks around 5 and 150 in the horizontal axis corresponds to the sum of the diagonal components of the density matrix. The result of the measurement is $\rho_{\uparrow\uparrow\uparrow\uparrow} + \rho_{\downarrow\downarrow\downarrow\downarrow} = 0.69 \pm 0.02$.

Figure 3b shows a sinusoidal oscillation of the parity (circles) and a fit with a sinusoidal function (solid curve). The amplitude of the sinusoidal oscillation in the parity corresponds to the sum of the off-diagonal components of the density matrix, $\rho_{\uparrow\uparrow\downarrow\downarrow}$ and $\rho_{\downarrow\downarrow\uparrow\uparrow}$. The value for the experimental parity result is $\rho_{\uparrow\uparrow\downarrow\downarrow} + \rho_{\downarrow\downarrow\uparrow\uparrow} = 0.35 \pm 0.05$. Therefore, the fidelity for the $(|\uparrow\rangle_1|\uparrow\rangle_2 + |\downarrow\rangle_1|\downarrow\rangle_2)/\sqrt{2}$ state is 0.52 ± 0.03 (see Methods; the confidence intervals are for a 68% confidence level). Thus, we cannot state that this value clears the 0.5 threshold[21] when the confidence interval is considered. However, if we consider the reduction of the fidelity due to the imperfections in the analysis pulses and make a corresponding correction, the fidelity is estimated to be 0.74 ± 0.05 (see equation (19) in Methods). Thus, we speculate that an entangled phonon state is successfully generated in the experiment.

We have demonstrated two-phonon interference in a trapped-ion system. This is an essential step towards the realization of boson sampling[7,22–25] with trapped ions. Recent advancements in the field of cavity optomechanics[26] have enabled control of the vibrational motion of micro- and nanoscopic mechanical systems at the level of a single vibrational quantum. In those systems, coupling of multiple phonon modes may become possible[26], thereby enabling multi-mode phonon dynamics, as demonstrated here.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 30 June; accepted 8 September 2015.

1. Hong, C. K., Ou, Z. Y. & Mandel, L. Measurement of subpicosecond time intervals between 2 photons by interference. *Phys. Rev. Lett.* **59,** 2044–2046 (1987).
2. Brown, K. R. *et al.* Coupled quantized mechanical oscillators. *Nature* **471,** 196–199 (2011).
3. Harlander, M., Lechner, R., Brownnutt, M., Blatt, R. & Hänsel, W. Trapped-ion antennae for the transmission of quantum information. *Nature* **471,** 200–203 (2011).
4. Haze, S., Tateishi, Y., Noguchi, A., Toyoda, K. & Urabe, S. Observation of phonon hopping in radial vibrational modes of trapped ions. *Phys. Rev. A* **85,** 031401(R) (2012).
5. Porras, D. & Cirac, J. I. Bose-Einstein condensation and strong-correlation behavior of phonons in ion traps. *Phys. Rev. Lett.* **93,** 263602 (2004).
6. Toyoda, K., Matsuno, Y., Noguchi, A., Haze, S. & Urabe, S. Experimental realization of a quantum phase transition of polaritonic excitations. *Phys. Rev. Lett.* **111,** 160501 (2013).
7. Shen, C., Zhang, Z. & Duan, L. M. Scalable implementation of boson sampling with trapped ions. *Phys. Rev. Lett.* **112,** 050504 (2014).
8. Santori, C., Fattal, D., Vuckovic, J., Solomon, G. S. & Yamamoto, Y. Indistinguishable photons from a single-photon device. *Nature* **419,** 594–597 (2002).
9. Beugnon, J. *et al.* Quantum interference between two single photons emitted by independently trapped atoms. *Nature* **440,** 779–782 (2006).
10. Kaltenbaek, R., Blauensteiner, B., Zukowski, M., Aspelmeyer, M. & Zeilinger, A. Experimental interference of independent photons. *Phys. Rev. Lett.* **96,** 240502 (2006).
11. Maunz, P. *et al.* Quantum interference of photon pairs from two remote trapped atomic ions. *Nature Phys.* **3,** 538–541 (2007).
12. Lang, C. *et al.* Correlations, indistinguishability and entanglement in Hong–Ou–Mandel experiments at microwave frequencies. *Nature Phys.* **9,** 345–348 (2013).
13. Kaufman, A. M. *et al.* Two-particle quantum interference in tunnel-coupled optical tweezers. *Science* **345,** 306–309 (2014).
14. Lopes, R. *et al.* Atomic Hong–Ou–Mandel experiment. *Nature* **520,** 66–68 (2015).
15. Liu, R. C., Odom, B., Yamamoto, Y. & Tarucha, S. Quantum interference in electron collision. *Nature* **391,** 263–265 (1998).

16. Bocquillon, E. *et al.* Coherence and indistinguishability of single electrons emitted by independent sources. *Science* **339,** 1054–1057 (2013).
17. Knill, E., Laflamme, R. & Milburn, G. J. A scheme for efficient quantum computation with linear optics. *Nature* **409,** 46–52 (2001).
18. Moehring, D. L. *et al.* Entanglement of single-atom quantum bits at a distance. *Nature* **449,** 68–71 (2007).
19. Cirac, J. I. & Zoller, P. Quantum computations with cold trapped ions. *Phys. Rev. Lett.* **74,** 4091–4094 (1995).
20. Mølmer, K. & Sørensen, A. Multiparticle entanglement of hot trapped ions. *Phys. Rev. Lett.* **82,** 1835–1838 (1999).
21. Sackett, C. A. *et al.* Experimental entanglement of four particles. *Nature* **404,** 256–259 (2000).
22. Broome, M. A. *et al.* Photonic boson sampling in a tunable circuit. *Science* **339,** 794–798 (2013).
23. Spring, J. B. *et al.* Boson sampling on a photonic chip. *Science* **339,** 798–801 (2013).
24. Tillmann, M. *et al.* Experimental boson sampling. *Nature Photon.* **7,** 540–544 (2013).
25. Crespi, A. *et al.* Integrated multimode interferometers with arbitrary designs for photonic boson sampling. *Nature Photon.* **7,** 545–549 (2013).
26. Aspelmeyer, M., Kippenberg, T. J. & Marquard, F. Cavity optomechanics. *Rev. Mod. Phys.* **86,** 1391–1452 (2014).

Acknowledgements We thank Y. Yamamoto for suggestions made at the early stages of this study, and K. Hayasaka for comments on the manuscript. This work was supported by JSPS KAKENHI, grant number 26400418.

Author Contributions S.U., K.T. and A.N. designed this study. A.N. and R.H. conducted the experiment. R.H. wrote an early version of the manuscript. K.T. revised and completed the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to K.T. (toyoda@ee.es.osaka-u.ac.jp).

METHODS

Hopping Hamiltonian and beam-splitter transformation. We assume that two ions are confined in a linear Paul trap. When the harmonic confinement in the radial directions is significantly stronger than the Coulomb forces (the condition for 'stiff modes'[5]), the radial phonons can be regarded as local phonons confined in each ion site. In this situation, the radial phonons behave as bosonic particles, their total number is conserved, and the Coulomb interaction induces hopping of the radial phonons between the two ion sites.

The Hamiltonian that governs the motion in one of the radial directions (which we refer to as the y direction here) can be expressed as

$$\hat{H}_y = \sum_{i=1,2} \hbar\left(\omega_y - \frac{\kappa}{2}\right)\hat{a}_i^\dagger \hat{a}_i + \frac{\hbar\kappa}{2}(\hat{a}_1\hat{a}_2^\dagger + \hat{a}_1^\dagger\hat{a}_2) \tag{4}$$

where ω_y is the oscillation frequency in the y direction and

$$\kappa = \frac{e^2}{4\pi\epsilon_0 d_0^3 M\omega_y} \tag{5}$$

is the phonon hopping rate[4] (ϵ_0 is the vacuum permittivity, $d_0 = |z_1 - z_2|$ is the inter-ion distance in the axial direction and M is the ion mass). $\hat{a}_i$ and $\hat{a}_i^\dagger$ are, respectively, the annihilation and creation operators of local phonons in the ith site. The first term on the right-hand side of equation (4) describes the harmonic oscillators associated with the ion sites, while the second term represents phonon hopping between the two sites. By moving to an interaction picture, where the trivial dynamics due to the first term are omitted, we obtain the interaction Hamiltonian

$$\hat{H}_I = \frac{\hbar\kappa}{2}(\hat{a}_1\hat{a}_2^\dagger + \hat{a}_1^\dagger\hat{a}_2) \tag{6}$$

The propagator for this Hamiltonian is

$$\hat{U}_I(t) = \exp\left(-\frac{i\hat{H}_I t}{\hbar}\right) = \exp\left[-\frac{i\kappa t}{2}(\hat{a}_1\hat{a}_2^\dagger + \hat{a}_1^\dagger\hat{a}_2)\right] \tag{7}$$

and the annihilation operator in the Heisenberg picture is

$$\begin{aligned}\hat{a}_i(t) &= \hat{U}_I^\dagger(t)\hat{a}_i(0)\hat{U}_I(t) \\ &= \hat{a}_i\cos\left(\frac{\kappa t}{2}\right) - i\hat{a}_j\sin\left(\frac{\kappa t}{2}\right), \quad (i,\ j=1,2;\ i\neq j)\end{aligned} \tag{8}$$

where

$$\hat{a}_i(0) = \hat{a}_i\,, \qquad (i,j=1,2) \tag{9}$$

For two ions, $\hat{H}_I$ couples $|n\rangle_1|n+1\rangle_2$ and $|n+1\rangle_1|n\rangle_2$, causing phonon energies to be exchanged between the ions at κ. Here, $|n\rangle_i$ represents the phonon Fock state of the ith ion with the quantum number n. The 50:50 beam-splitter transformation, which is routinely used in linear optics, can be described as

$$\hat{b}_1 = (\hat{a}_1 - i\hat{a}_2)/\sqrt{2} \tag{10}$$

$$\hat{b}_2 = (-i\hat{a}_1 + \hat{a}_2)/\sqrt{2} \tag{11}$$

where $\hat{b}_i$ ($i=1,2$) is the annihilation operator for each of the two beam-splitter outputs. $\hat{U}_I(t)$ with $t=\pi/2\kappa$ corresponds to this transformation and $\hat{a}_1(\pi/2\kappa)$ and $\hat{a}_2(\pi/2\kappa)$ are equal to $\hat{b}_1$ and $\hat{b}_2$ in equations (10) and (11), respectively. $\pi/2\kappa$ is equal to $T_{\rm hop}/4$, where $T_{\rm hop} = 2\pi/\kappa$ is the hopping period. With the transformation given above, an initial product state $|1\rangle_1|1\rangle_2$ is transformed to $-i(|2\rangle_1|0\rangle_2 + |0\rangle_1|2\rangle_2)/\sqrt{2}$.

In addition to the local-mode picture described immediately above, the two-phonon dynamics and the HOM effect can be understood in terms of the collective-mode picture. The hopping Hamiltonian, equation (6), can be diagonalized in the two-phonon subspace using the following three eigenkets in the collective-mode picture

$$|2_c0_r\rangle \equiv \frac{1}{\sqrt{2}}\hat{a}_c^{\dagger 2}|0\rangle_1|0\rangle_2 = \frac{1}{2}|2\rangle_1|0\rangle_2 + \frac{1}{\sqrt{2}}|1\rangle_1|1\rangle_2 + \frac{1}{2}|0\rangle_1|2\rangle_2 \tag{12}$$

$$|1_c1_r\rangle \equiv \hat{a}_r^\dagger\hat{a}_c^\dagger|0\rangle_1|0\rangle_2 = \frac{1}{\sqrt{2}}|2\rangle_1|0\rangle_2 - \frac{1}{\sqrt{2}}|0\rangle_1|2\rangle_2 \tag{13}$$

$$|0_c2_r\rangle \equiv \frac{1}{\sqrt{2}}\hat{a}_r^{\dagger 2}|0\rangle_1|0\rangle_2 = \frac{1}{2}|2\rangle_1|0\rangle_2 - \frac{1}{\sqrt{2}}|1\rangle_1|1\rangle_2 + \frac{1}{2}|0\rangle_1|2\rangle_2 \tag{14}$$

with eigenvalues $+\hbar\kappa$, 0 and $-\hbar\kappa$, respectively. Here

$$\hat{a}_c^\dagger = \frac{1}{\sqrt{2}}(\hat{a}_1^\dagger + \hat{a}_2^\dagger) \tag{15}$$

$$\hat{a}_r^\dagger = \frac{1}{\sqrt{2}}(\hat{a}_1^\dagger - \hat{a}_2^\dagger) \tag{16}$$

are the creation operators for the radial centre-of-mass (c.m.) and rocking modes, respectively. Using these eigenkets, an initial Fock state $|1\rangle_1|1\rangle_2$ can be expressed as

$$|1\rangle_1|1\rangle_2 = \frac{1}{\sqrt{2}}(|2_c0_r\rangle - |0_c2_r\rangle) \tag{17}$$

Under the dynamics due to the hopping Hamiltonian, this state alternates with the following state (apart from a phase factor) with period $\frac{2\pi}{2\kappa} = \frac{1}{2}T_{\rm hop}$

$$\frac{1}{\sqrt{2}}(|2\rangle_1|0\rangle_2 + |0\rangle_1|2\rangle_2) = \frac{1}{\sqrt{2}}(|2_c0_r\rangle + |0_c2_r\rangle) \tag{18}$$

These dynamics cause an oscillation of the phonon coincidence between the two sites, which can be observed in experiments.

Experimental procedure. The 729-nm laser propagates in a direction that creates angles of 45°, 45° and 90° with the x, y and z directions, respectively. Sideband cooling is performed for both the x and y directions, where x is the other of the two radial directions that is orthogonal to y. The axial direction, z, is cooled only by Doppler cooling. There are two collective modes in both the x and y directions, namely, the c.m. (in-phase) mode and the rocking (out-of-phase) mode. Their frequency separation is of the order of $\kappa/2\pi \approx 2$ kHz, and here, for both x and y, the c.m. and rocking modes are cooled at the same time. The average quantum numbers in the y direction after the sideband cooling are $(\bar{n}_{y,\rm c.m.}, \bar{n}_{y,\rm rock}) = (0.04 \pm 0.07, 0.03 \pm 0.12)$. For the x direction, $\bar{n}_{x,\rm c.m.}$ and $\bar{n}_{x,\rm rock}$ are estimated to be < 0.7 and < 0.2, respectively.

Individual observation of the two ions is enabled by illuminating them with unequal intensities; this is achieved by displacing the centre of the waist of the 397-nm beam from the midpoint of the inter-ion distance, so that their fluorescence levels differ.

Numerical simulation and fidelity of sideband Rabi pulses. We performed a numerical simulation of the whole dynamics of the two-ion system irradiated with the lasers, in order to confirm the experimental results and to support the fidelity analysis. We used a Liouville equation with Lindblad-type relaxation terms. The parameters used for the sideband Rabi rotations in the simulation were a frequency of 26.3 kHz and an e^{-1} decay time of 100 μs. We assumed the latter to be due to a pure phase relaxation. These two quantities were estimated from experimental results for sideband Rabi oscillations. Finally, κ was assumed to be $2\pi \times 2.05$ kHz. The reason for the relatively fast decay time of the sideband Rabi oscillation (100 μs) is currently unknown. The possible causes are: phase jitter of the excitation pulse due to beam jitter, fluctuations of AC Stark shifts due to the 729-nm laser, and relaxation of the motional coherence due to other motional modes.

We estimated the expected fidelity based on this simulation. The sum of the diagonal components of the density matrix, the sum of its off-diagonal components, and the fidelity were found to be, respectively, $(0.742 \pm 0.000, 0.407 \pm 0.014, 0.575 \pm 0.007)$. As stated above, this was obtained while assuming relatively fast relaxation in the sideband Rabi oscillations. The resultant imperfections and non-fidelities are relatively large. If we assume perfect fidelity for the analysis pulses (the red-sideband π pulse on $|g,2\rangle_i - |e,1\rangle_i$, the blue-sideband $\pi/2$ pulse on $|g,0\rangle_i - |e,1\rangle_i$ and, additionally, the blue-sideband $\pi/2$ pulse with varying phase on $|g,0\rangle_i - |e,1\rangle_i$ in the case of parity analysis), the quantities quoted above are found to be $(0.885 \pm 0.000, 0.753 \pm 0.032, 0.819 \pm 0.016)$, respectively.

From these two cases, we estimated the ratios of the reduction of the diagonal and off-diagonal components due to the imperfections in the sideband Rabi rotations, which were found to be 0.838 ± 0 and 0.541 ± 0.030, respectively. If these ratios are used to divide the experimental values (0.691 ± 0.020 and 0.352 ± 0.047, respectively), the diagonal and off-diagonal components expected in the case of ideal analysis pulses without imperfection are 0.825 ± 0.024 and 0.651 ± 0.094, respectively. The fidelity in this case is then estimated to be 0.738 ± 0.048. This value well exceeds the threshold of 0.5 for the entangled states[21].

Imperfections in hopping and coincidence results. In the result shown in Fig. 2b, the maximum value of the phonon population does not reach 1. In addition, the results do not begin from this value at the origin of the horizontal axis. This behaviour would not be observed in an ideal case involving instant preparation and analysis of the phonon states with perfect fidelity. Instead, the phonon population in ion 1 would begin at the maximum value, which should be 1. The former

imperfection is due to the reduction of the fidelity in the sideband Rabi rotations (see the previous section), which causes imperfect preparation of the phonon Fock states. The latter imperfection is due to the non-negligible lengths of the pulses used for preparation and analysis compared with the hopping period. The blue-sideband π and quenching pulses used for the preparation have durations of ∼9 μs and ∼30 μs, respectively, and the red-sideband π pulse used for analysis (mapping) has a duration of ∼ 19 μs. In Fig. 2b, it can be seen that the time origin of the dynamics is shifted in the negative direction by ∼ 67 μs, which is not very different from the sum of the above three values.

The imperfections in the maximum value and the shift of the time origin in Fig. 2c can be explained in a similar manner to the case shown in Fig. 2b.

Fidelity of $(|2\rangle_1|0\rangle_2 + |0\rangle_1|2\rangle_2)/\sqrt{2}$ entangled state. We estimated the fidelity of the $(|2\rangle_1|0\rangle_2 + |0\rangle_1|2\rangle_2)/\sqrt{2}$ entangled state at $t \approx 56$ μs in the following manner. (1) A red-sideband π pulse that was resonant with the $|g,2\rangle_i \leftrightarrow |e,1\rangle_i$ transition was applied to $(|2\rangle_1|0\rangle_2 + |0\rangle_1|2\rangle_2)/\sqrt{2}$. Thus, this state was transferred to $(|e,1\rangle_1|g,0\rangle_2 + |g,0\rangle_1|e,1\rangle_2)/\sqrt{2}$. By rewriting $|e,1\rangle_i$ as $|\uparrow\rangle_i$ and $|g,0\rangle_i$ as $|\downarrow\rangle_i$, this state could be expressed as $(|\uparrow\rangle_1|\downarrow\rangle_2 + |\downarrow\rangle_1|\uparrow\rangle_2)/\sqrt{2}$ (2) A blue-sideband $\pi/2$ pulse that was resonant with the $|\uparrow\rangle_i \leftrightarrow |\downarrow\rangle_i$ transition was applied to $(|\uparrow\rangle_1|\downarrow\rangle_2 + |\downarrow\rangle_1|\uparrow\rangle_2)/\sqrt{2}$. Thus, this state was transferred to $(|\uparrow\rangle_1|\uparrow\rangle_2 + |\downarrow\rangle_1|\downarrow\rangle_2)/\sqrt{2}$. (3) In order to estimate the diagonal components of the density matrix, $\rho_{\downarrow\downarrow\downarrow\downarrow}$ and $\rho_{\uparrow\uparrow\uparrow\uparrow}$, the fluorescence of the two ions was recorded by illuminating the system with the 397-nm laser at this point. (4) To measure the non-diagonal components of the density matrix, $\rho_{\downarrow\downarrow\uparrow\uparrow}$ and $\rho_{\uparrow\uparrow\downarrow\downarrow}$, we performed a parity measurement. A 729-nm blue-sideband $\pi/2$ pulse with varying phase was applied to $(|\uparrow\rangle_1|\uparrow\rangle_2 + |\downarrow\rangle_1|\downarrow\rangle_2)/\sqrt{2}$. and the fluorescence of the two ions was recorded.

We obtained the fidelity of the state $(|2\rangle_1|0\rangle_2 + |0\rangle_1|2\rangle_2)/\sqrt{2}$ using the relation

$$F \equiv |\langle\psi|\rho_{\mathrm{exp}}|\psi\rangle|^2 = \frac{1}{2}(\rho_{\downarrow\downarrow\downarrow\downarrow} + \rho_{\uparrow\uparrow\uparrow\uparrow} + \rho_{\downarrow\downarrow\uparrow\uparrow} + \rho_{\uparrow\uparrow\downarrow\downarrow}) \tag{19}$$

If this value exceeds 0.5, we regard the generated state as an entangled state[21].

LETTER

doi:10.1038/nature15746

An aqueous, polymer-based redox-flow battery using non-corrosive, safe, and low-cost materials

Tobias Janoschka[1,2], Norbert Martin[3], Udo Martin[3], Christian Friebe[1,2], Sabine Morgenstern[1,2], Hannes Hiller[1,2], Martin D. Hager[1,2] & Ulrich S. Schubert[1,2]

For renewable energy sources such as solar, wind, and hydroelectric to be effectively used in the grid of the future, flexible and scalable energy-storage solutions are necessary to mitigate output fluctuations[1]. Redox-flow batteries (RFBs) were first built in the 1940s[2] and are considered a promising large-scale energy-storage technology[1,3,4]. A limited number of redox-active materials[4,5–10]—mainly metal salts, corrosive halogens, and low-molar-mass organic compounds—have been investigated as active materials, and only a few membrane materials[3,5,11–14], such as Nafion, have been considered for RFBs. However, for systems that are intended for both domestic and large-scale use, safety and cost must be taken into account as well as energy density and capacity, particularly regarding long-term access to metal resources, which places limits on the lithium-ion-based and vanadium-based RFB development[15,16]. Here we describe an affordable, safe, and scalable battery system, which uses organic polymers as the charge-storage material in combination with inexpensive dialysis membranes, which separate the anode and the cathode by the retention of the non-metallic, active (macro-molecular) species, and an aqueous sodium chloride solution as the electrolyte. This water- and polymer-based RFB has an energy density of 10 watt hours per litre, current densities of up to 100 milliamperes per square centimetre, and stable long-term cycling capability. The polymer-based RFB we present uses an environmentally benign sodium chloride solution and cheap, commercially available filter membranes instead of highly corrosive acid electrolytes and expensive membrane materials.

A growing interest in organic redox-active materials has been observed over the last few years (Extended Data Table 1). Semi-organic systems include TEMPO (2,2,6,6-tetramethylpiperidinyloxyl)/lithium[8], anthraquinone/lithium[17], and viologen/lithium[18]. Despite high cell voltages and good theoretical capacities, the power capability of these systems is limited, owing to low ion mobility in the organic solvents used. This drawback can be overcome by using acid-based, aqueous electrolytes as suggested for a bromine–anthraquinone cell[6]. However, bromine and highly acidic electrolytes represent a substantial risk and a challenge to all applied system components (for example, corrosion of pumps, pipes, and storage tanks), and so interest in all-organic RFBs arose[7,9,19]. Yet these systems have low capacities (at most 5 W h l^{-1}), poor current densities (at most 10 mA cm^{-2}), and poor cycling stabilities (below 30 cycles).

Virtually all realized RFBs are made of two electrolyte circuits separated by an ion-selective membrane. Considerable effort has been invested to develop membrane materials that show a low area resistivity, are chemically resistant to acidic electrolytes, and are highly selective to prevent cross-contamination of the electrolytes. Nevertheless, "the membrane has been identified as one of the main obstacles in the commercialisation of many redox flow cells"[11].

Perfluorinated ion-exchange membranes are commonly used, because they are robust and withstand a highly oxidative and corrosive environment. However, Nafion—the most commonly used material—accounts for almost 40% of the cost of the reaction cell. Alternative membrane materials include microporous membranes such as (filled and/or modified) Daramic or Celgard[3,5,11–14]; the latter has also been tested with organic polymer solutions. Celgard (pore radius of 14–21 nm) targets a pore-size exclusion effect (steric hindrance), but only works with polymers of very high molar mass[18,20]. Several attempts to use nanofiltration membranes (pore radii in the 1-nm range) in vanadium-based RFBs showed that it is possible to achieve vanadium/proton selectivity by means of pore-size exclusion. However, the corrosive electrolyte is highly demanding in terms of chemical membrane stability, and large-scale applications require inexpensive, easy-to-manufacture materials[21–23].

The aforementioned challenges illustrate the need for a battery system that combines a water-based electrolyte with an organic redox-active material and a suitable low-cost membrane. Here we propose a new RFB design that fulfils these demands by using (1) organic polymers as the redox-active species, (2) an aqueous sodium chloride solution as the electrolyte, and (3) simple dialysis membranes (Fig. 1).

Dialysis membranes, which can retain macromolecules of high molar mass while allowing small ions to pass regardless of their charge, are affordable and widely used—from laboratory-scale experiments to industrial water-treatment facilities[24]. For the proposed polymer-based RFB, we chose a cellulose-based dialysis membrane with a molecular-weight cut-off (MWCO; indicating the lowest retained molar mass) of 6,000 g mol^{-1} and an aqueous sodium chloride solution as the supporting electrolyte. Both components were selected because of their compatibility with the chosen redox-active polymers. A large number of redox-active polymers have been studied for use in solid-state batteries in the past[25,26], but only compounds that show stable redox behaviour in water are suitable for the proposed RFB system. Therefore, extensive screening of potential polymers was performed.

The optimized polymers consist of two components: a redox-active moiety and a unit enhancing water solubility to prevent precipitation in all used redox states. The cathode material contains the TEMPO radical as the redox-active moiety, while the anode material uses a 4,4′-bipyridine derivative (viologen). The water-solubility of both polymers is enhanced by a quaternary ammonium cation moiety. The cathode material was prepared by free radical copolymerization of 2,2,6,6-tetramethylpiperidin-4-yl-methacrylate **1** and amine **2** (for synthetic details see Extended Data Fig. 1). Subsequent oxidation with H_2O_2/Na_2WO_4 yielded the desired polymer **P1** (Fig. 1a). The anode material **P2** (Fig. 1a) is obtained by copolymerization of 4-vinylbenzyl chloride **3** and the amine **4** (Extended Data Fig. 1), followed by polymer-analogous functionalization with *N*-methyl-bipyridinium iodide and an ion exchange to chloride. Both materials were prepared at kilogram scale. A modifier (2-mercaptoethanol) was used in the preparation of **P1** to guarantee a low molar mass (M_n) of about 20,000 g mol^{-1} and a low dispersity; this was not necessary in the second polymerization process. The molar-mass target was chosen to achieve both a good

[1]Laboratory of Organic and Macromolecular Chemistry (IOMC), Friedrich Schiller University Jena, Humboldtstrasse 10, 07743 Jena, Germany. [2]Center for Energy and Environmental Chemistry Jena (CEEC Jena), Friedrich Schiller University Jena, Philosophenweg 7a, 07743 Jena, Germany. [3]JenaBatteries GmbH, Botzstrasse 5, 07743 Jena, Germany.

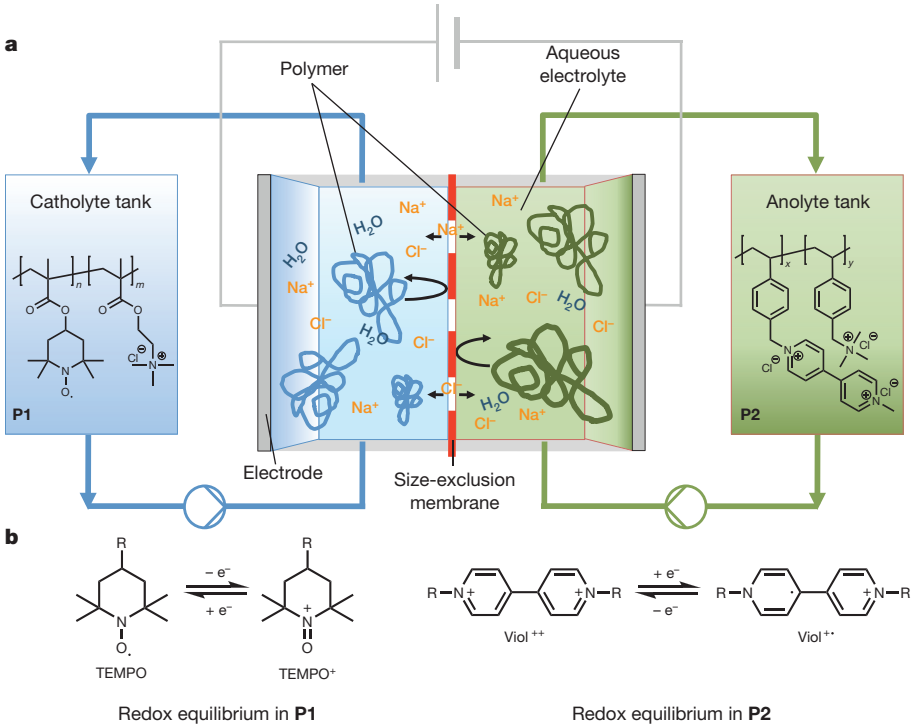


Figure 1 | Working principle of a polymer-based RFB. a, Schematic representation of a polymer-based RFB consisting of an electrochemical cell (which determines the power density) and two electrolyte reservoirs (which determine the storage capacity). The anolyte and catholyte cycle are separated by a semipermeable size-exclusion membrane, which retains the redox-active macromolecules while allowing small salt ions to pass. During the charging/discharging process, a solution of the redox-active polymers **P1** and **P2** is continuously transported from the electrolyte reservoirs to the electrochemical cell, where the redox reactions take place. **b**, Fundamental electrode reactions of **P1** (TEMPO radical) and **P2** (viologen). Structural details of compounds shown in this figure are available as Supplementary Information.

retention of the polymer by the dialysis membrane and a dynamic viscosity as low as possible.

Rheological investigations of the catholyte and anolyte, each with a capacity of about 10 A h l^{-1}, revealed Newtonian behaviour and apparent viscosities of 17 mPa s (**P1**) and 5 mPa s (**P2**) in the shear-rate range that is typically attributed to pipe flow (Extended Data Fig. 2). Consequently, in contrast to RFBs that use a polymer of high molar mass and microporous Celgard membranes[18], the energy required to pump the electrolyte is kept to a minimum, which facilitates efficient transport of the solutions through the reaction cell.

The performance of an RFB is strongly influenced by the quality of the membrane. It has to retain the redox-active species, thus preventing internal short-circuits and self-discharge processes, while facilitating the transport of ions, which are necessary to sustain electro-neutrality. The salt permeability (P_S) for sodium chloride through the studied cellulose-based dialysis membrane—the thickness-normalized diffusion coefficient—was found to be $(9.3 \pm 0.1) \times 10^{-5}$ cm s^{-1} (the uncertainty here and elsewhere corresponds to error propagation of linear-regression-analysis error; Extended Data Fig. 4). This leads to a low area resistance (R) of the membrane in aqueous sodium chloride solution of $1.14 \pm 0.03\ \Omega\ cm^2$, which is in the range of Nafion and enables good cell performance[10]. The redox-active polymers, which have a hydrodynamic radius of around 2 nm (determined using dynamic light scattering), are effectively retained by the dialysis membrane, which has an estimated pore size <1 nm. Minimum membrane selectivities (S_{min}) for sodium chloride over the redox-active polymers of 290 (**P1**) and 2,830 (**P2**) were obtained, which are much higher than the membrane selectivities of Nafion[11], and those of nanofiltration[21,22] and microporous membranes[18]. In addition, the retention was tested under real-life conditions in an RFB test cell. Cyclic-voltammetry measurements of samples taken from the anolyte and the catholyte solutions after 10,000 charging/discharging cycles show no detectable amounts of polymer **P1** transferred from one cell compartment to the other (with a detection limit of approximately 2 μg ml^{-1}) and only traces of **P2** (Extended Data Fig. 5). This finding is supported by the consistently high coulombic efficiency of 99%.

The redox properties of the TEMPO/viologen redox pair were studied via cyclic voltammetry (Extended Data Fig. 6). The basic electrode reactions are displayed in Fig. 1b. Upon charging, the TEMPO radical is oxidized, forming an oxammonium cation ($TEMPO^+$), while the divalent viologen cation ($Viol^{++}$) is reduced to a monovalent radical cation ($Viol^{+\bullet}$). Cyclic voltammetry of **P1** reveals a reversible redox wave at 0.7 V (versus the Ag/AgCl reference electrode), in accordance with literature values for the TEMPO radical[25]. The viologen-containing polymer **P2** shows quasi-reversible redox reactions at −0.4 V and −0.8 V. Because the second reaction yields a neutral, water-insoluble species ($Viol^0$), only the first step was studied in detail:

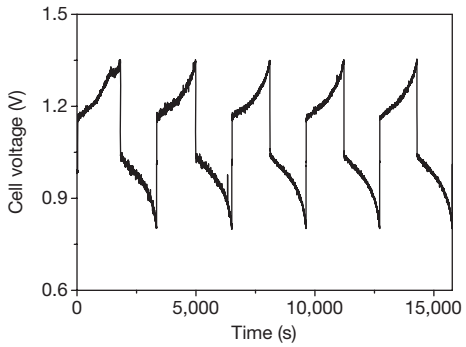


Figure 2 | Charge/discharge behaviour. A representative cell voltage profile of a pumped 5-cm^2 test cell during constant-current cycling at 40 mA cm^{-2} with 10 ml of **P1** and 15 ml of **P2** solution (charge storage capacity adjusted to 10 A h l^{-1} in aqueous NaCl solution (2 mol l^{-1}), 25 °C).

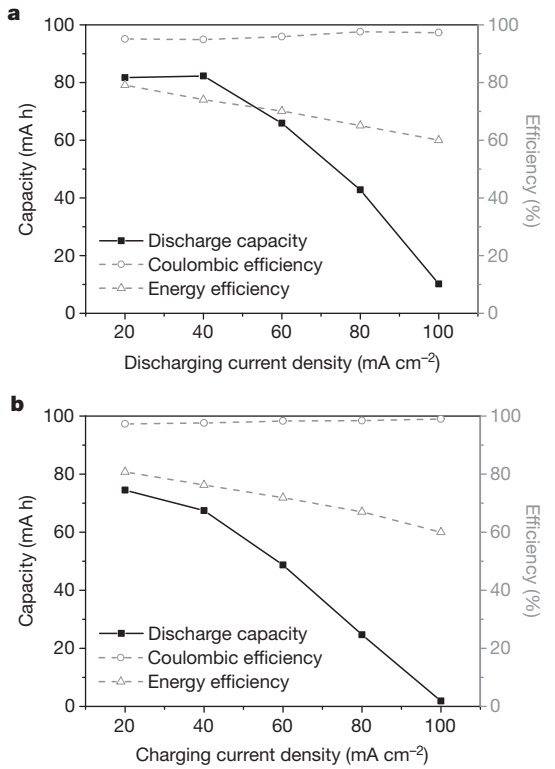


Figure 3 | Electric performance of the polymer-based RFB cell. **a**, **b**, The capacity, coulombic efficiency, and energy efficiency of a pumped 5-cm^2 test cell (10 ml of **P1** and 15 ml of **P2** aqueous NaCl solution (2 mol l^{-1}); storage capacity adjusted to 10 A h l^{-1}, 25 °C) as a function of discharging current density (**a**; charging at 40 mA cm^{-2}) and charging current density (**b**; discharging at 40 mA cm^{-2}).

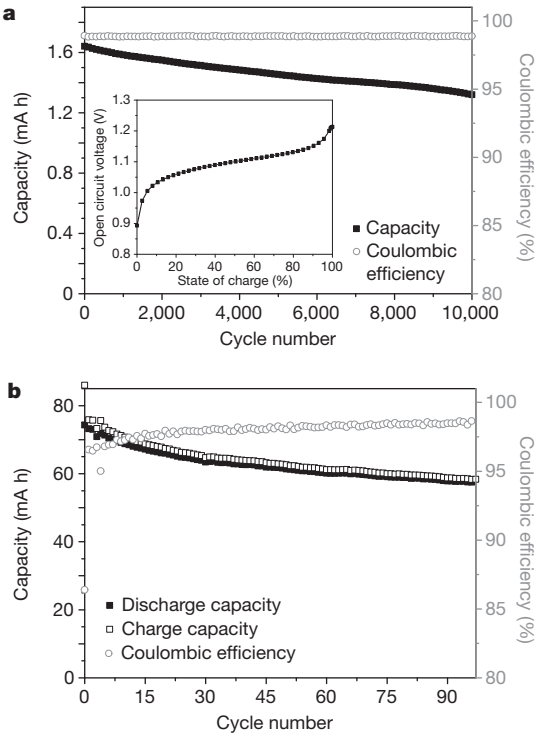


Figure 4 | Cycling stability of the polymer-based RFB. **a**, The long-term stability of the polymer-based electrolytes was studied by repeated charge/discharge cycling over 10,000 cycles at 20 mA cm^{-2} in an unpumped test cell. Inset, the open-circuit voltage of a polymer-based RFB as a function of the state of charge (static 5-cm^2 test cell, **P1** (storage capacity of 2 A h l^{-1}) and **P2** (storage capacity of 4 A h l^{-1}) in aqueous NaCl solution (2 mol l^{-1}), 25 °C). **b**, A pumped cell was repeatedly cycled at 40 mA cm^{-2} (10 ml of **P1** and 15 ml of **P2** in aqueous NaCl solution (2 mol l^{-1}); storage capacity adjusted to 10 A h l^{-1}, 25 °C).

cyclic voltammetry reveals two waves at −0.40 V and −0.53 V for the reduction process, and a sharp peak at −0.45 V and a broader signal at −0.38 V for the re-oxidation. This signal split can be attributed to a reversible intramolecular association/dimer formation between two viologen radical cations[27,28], which is supported by ultraviolet–visible-spectroelectrochemical studies: upon reduction, a set of absorption bands that are characteristic of the formation of viologen radical dimers[27] arises (Extended Data Fig. 7). Additionally, applying a re-oxidizing potential restores the initial spectrum, indicating reversibility of the redox process.

We conducted rotating-disk-electrode (RDE) voltammetry to obtain further kinetic data (Extended Data Figs 8 and 9); see Methods for details of the subsequent analysis. Levich analysis of the voltammograms, obtained for a variety of rotation speeds, yields diffusion coefficients (D) of $(7.0 \pm 0.5) \times 10^{-8}$ cm^2 s^{-1} and $(7.6 \pm 0.9) \times 10^{-7}$ cm^2 s^{-1} for the polymers **P1** and **P2**, respectively. Subsequent Koutecký–Levich analysis for **P1** provides mass-transport-independent currents, which are fitted to the Butler–Volmer equation to obtain an electron-transfer rate constant (k^0) of $(4.5 \pm 0.1) \times 10^{-4}$ cm s^{-1}. Tafel analysis determines the transfer coefficient (α) for **P1** to be 0.68 ± 0.03, which is fairly close to 0.5, the value for an ideally reversible redox reaction. For **P2**, RDE voltammetry yielded analysable data only for high rotation rates, and a two-step mechanism is assumed for the first reduction process on the basis of the obtained voltammograms. Appropriate analysis yields $k^0 = (9 \pm 2) \times 10^{-5}$ cm s^{-1}. The standard electron-transfer rates for **P1** and **P2** are in the range of common small-molecule RFB redox-active materials[29].

A battery was built using aqueous solutions of the redox-active polymers **P1** and **P2**, a cellulose-based dialysis membrane (5 cm^2 active area), and sodium chloride as supporting electrolyte. With its MWCO of 6,000 g mol^{-1}, the membrane can effectively retain both polymers (with molar masses three times larger than the MWCO). The cell provides an open-circuit voltage of 1.1 V and can be safely charged and discharged within a voltage window of 0.80–1.35 V; we did not observe evolution of oxygen, hydrogen, or chlorine. Because the charging process is accompanied by a strong colour shift from orange to yellow (**P1**) and ochre to blue (**P2**), solution colour represents a simple indication of the state of charge of the battery (Extended Data Fig. 3). Representative charging/discharging curves for constant-current cycling at 40 mA cm^{-2} are displayed in Fig. 2. The cell can be charged and discharged within the chosen 'real-life' voltage window at current densities of up to 40 mA cm^{-2}, while retaining most of its initial capacity, and achieving an energy efficiency between 75% and 80% (Fig. 3). Pulse current densities of up to 100 mA cm^{-2} are possible. At a theoretical capacity of 10 A h l^{-1}, a material utilization of up to 82% was observed, which corresponds to energy densities of 10.8 W h l^{-1} (charging) and 8.0 W h l^{-1} (discharging). The observed performance approaches conventional vanadium-based RFBs and surpasses all-organic RFBs that use 'small' redox-active molecules in combination with conventional membranes[7,9,10,19].

Incremented charge/discharge experiments yielded a relatively flat, sigmoid open-circuit-voltage curve, which indicates a stable voltage between 10% and 90% state of charge and allows us to monitor the battery by measuring the open-circuit voltage. Cycling studies at constant current revealed good stability of the developed polymer-based RFB in comparison to other organic RFBs presented in the literature (Extended Data Table 1). Even after extended long-term tests of 10,000 cycles in a static, unpumped cell, 80% of the initial capacity was

retained (at 20 mA cm^{-2}). Because mass transport relies solely on diffusion in this set-up, the rapid charging allowed for a material utilization of only 41%. In a pumped cell, higher states of charge were achieved at 40 mA cm^{-2} with a material utilization of 75%. A faster capacity fade caused by a side-reaction can be observed. This might be induced by oxygen, which slowly enters the electrolyte as a result of mechanical abrasion of the tubes in the peristaltic pump causing oxidation of the viologen radical cation Viol$^{+\bullet}$ (Fig. 4).

By combining simple dialysis membranes, which are only 5% to 10% of the cost of Nafion, and safe polymer-based aqueous electrolytes, we designed an affordable RFB concept. We expect that further performance improvements can be achieved by optimizing the process parameters, hardware engineering, and polymer synthesis. Limitations resulting from the viscosity of the electrolyte solutions could be overcome by substituting linear polymers for (hyper-)branched polymers, the latter generally having lower viscosities at higher concentrations enabling increased energy densities. Besides the TEMPO and the viologen moieties, a large number of potential organic redox-active units might help to further boost cell voltage and cycling stability of future polymer-based RFBs. In any case, the presented work lays the foundation for a new battery principle, which could lead to the production of economical energy-storage devices that use safe, metal-free, and all-organic raw materials.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 31 March; accepted 29 September 2015.
Published online 21 October 2015.

1. Dunn, B., Kamath, H. & Tarascon, J.-M. Electrical energy storage for the grid: a battery of choices. *Science* **334,** 928–935 (2011).
2. Kangro, W. Verfahren zur Speicherung von elektrischer Energie. German patent DE 914 264 (1949).
3. Yang, Z. *et al.* Electrochemical energy storage for green grid. *Chem. Rev.* **111,** 3577–3613 (2011).
4. Alotto, P., Guarnieri, M. & Moro, F. Redox flow batteries for the storage of renewable energy: a review. *Renew. Sustain. Energy Rev.* **29,** 325–335 (2014).
5. Wang, W. *et al.* Recent progress in redox flow battery research and development. *Adv. Funct. Mater.* **23,** 970–986 (2013).
6. Huskinson, B. *et al.* A metal-free organic–inorganic aqueous flow battery. *Nature* **505,** 195–198 (2014).
7. Yang, B., Hoober-Burkhardt, L., Wang, F., Surya Prakash, G. K. & Narayanan, S. R. An inexpensive aqueous flow battery for large-scale electrical energy storage based on water-soluble organic redox couples. *J. Electrochem. Soc.* **161,** A1371–A1380 (2014).
8. Wei, X. *et al.* TEMPO-based catholyte for high-energy density nonaqueous redox flow batteries. *Adv. Mater.* **26,** 7649–7653 (2014).
9. Brushett, F. R., Vaughey, J. T. & Jansen, A. N. An all-organic non-aqueous lithium-ion redox flow battery. *Adv. Energy Mater.* **2,** 1390–1396 (2012).
10. Leung, P. *et al.* Progress in redox flow batteries, remaining challenges and their applications in energy storage. *RSC Adv.* **2,** 10125–10156 (2012).
11. Prifti, H., Parasuraman, A., Winardi, S., Lim, T. M. & Skyllas-Kazacos, M. Membranes for redox flow battery applications. *Membranes* **2,** 275–306 (2012).
12. Li, X., Zhang, H., Mai, Z., Zhang, H. & Vankelecom, I. Ion exchange membranes for vanadium redox flow battery (VRB) applications. *Energy Environ. Sci.* **4,** 1147–1160 (2011).
13. Shin, S.-H., Yun, S.-H. & Moon, S.-H. A review of current developments in non-aqueous redox flow batteries: characterization of their membranes for design perspective. *RSC Adv.* **3,** 9095–9116 (2013).
14. Schwenzer, B. *et al.* Membrane development for vanadium redox flow batteries. *ChemSusChem* **4,** 1388–1406 (2011).
15. Barnhart, C. J. & Benson, S. M. On the importance of reducing the energetic and material demands of electrical energy storage. *Energy Environ. Sci.* **6,** 1083–1092 (2013).
16. Armand, M. & Tarascon, J. M. Building better batteries. *Nature* **451,** 652–657 (2008).
17. Wang, W. *et al.* Anthraquinone with tailored structure for a nonaqueous metal–organic redox flow battery. *Chem. Commun.* **48,** 6669–6671 (2012).
18. Nagarjuna, G. *et al.* Impact of redox-active polymer molecular weight on the electrochemical properties and transport across porous separators in nonaqueous solvents. *J. Am. Chem. Soc.* **136,** 16309–16316 (2014).
19. Li, Z. *et al.* Electrochemical properties of an all-organic redox flow battery using 2,2,6,6-tetramethyl-1-piperidinyloxy and N-methylphthalimide. *Electrochem. Solid-State Lett.* **14,** A171–A173 (2011).
20. Sukegawa, T., Masuko, I., Oyaizu, K. & Nishide, H. Expanding the dimensionality of polymers populated with organic robust radicals toward flow cell application: synthesis of TEMPO-crowded bottlebrush polymers using anionic polymerization and ROMP. *Macromolecules* **47,** 8611–8617 (2014).
21. Zhang, H., Zhang, H., Li, X., Mai, Z. & Zhang, J. Nanofiltration (NF) membranes: the next generation separators for all vanadium redox flow batteries (VRBs)? *Energy Environ. Sci.* **4,** 1676–1679 (2011).
22. Xi, X. *et al.* Solvent responsive silica composite nanofiltration membrane with controlled pores and improved ion selectivity for vanadium flow battery application. *J. Power Sources* **274,** 1126–1134 (2015).
23. Zhou, X., Zhao, T. S., An, L., Wei, L. & Zhang, C. The use of polybenzimidazole membranes in vanadium redox flow batteries leading to increased coulombic efficiency and cycling performance. *Electrochim. Acta* **153,** 492–498 (2015).
24. Ulbricht, M. Advanced functional polymer membranes. *Polymer* **47,** 2217–2262 (2006).
25. Janoschka, T., Hager, M. D. & Schubert, U. S. Powering up the future: radical polymers for battery applications. *Adv. Mater.* **24,** 6397–6409 (2012).
26. Nishide, H., Koshika, K. & Oyaizu, K. Environmentally benign batteries based on organic radical polymers. *Pure Appl. Chem.* **81,** 1961–1970 (2009).
27. Imabayashi, S.-I., Kitamura, N., Tazuke, S. & Tokuda, K. Substituent effects on electrochemical reduction of viologen dimer and trimer with ethylene spacer. *J. Electroanal. Chem.* **239,** 397–403 (1988).
28. Imabayashi, S.-I., Kitamura, N., Tazuke, S. & Tokuda, K. The role of intramolecular association in the electrochemical reduction of viologen dimers and trimers. *J. Electroanal. Chem.* **243,** 143–160 (1988).
29. Weber, A. Z. *et al.* Redox flow batteries: a review. *J. Appl. Electrochem.* **41,** 1137–1164 (2011).

Supplementary Information is available in the online version of the paper.

Acknowledgements We acknowledge the European Regional Development Fund for Thuringia (EFRE), the Thüringer Aufbaubank (TAB), the Thuringian Ministry for Economic Affairs, Science and Digital Society (TMWWdG), and the Fonds der Chemischen Industrie for financial support. We thank J. Stammer, C. Oder, K. Wolkersdörfer, C. Stolze, C. Schmerbauch, B. Häupler, M. Wagner, T. Buś, A. Ignaszak, and F. Schacher for their assistance and comments.

Author Contributions T.J., M.D.H., and U.S.S. conceived the studies. N.M., C.F., and T.J. contributed to performing all electrochemical experiments and interpreting the results. S.M. and H.H. performed synthesis under the supervision of T.J. The test cell was designed by U.M. All authors discussed the results and commented on the manuscript. T.J., C.F., M.D.H., and U.S.S wrote the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details are available in the online version of the paper. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to U.S.S. (ulrich.schubert@uni-jena.de).

METHODS

Instruments and reagents. All reagents were bought from TCI, Sigma-Aldrich, and AlfaAesar, and used as received without further purification. *N*-Methyl-4,4′-bipyridinium iodide was synthesized according to literature procedures using the crude product for further reaction steps[30]. The polymerization reactions were carried out in an inert argon atmosphere. ^{1}H NMR spectra were obtained on a FOURIER 300 (Bruker), ATR-IR (infrared attenuated total reflection) spectra on an IRAffinity-1 (Shimadzu), and X-band EPR (electron paramagnetic resonance) spectra on an EMXmicro CW-EPR spectrometer (Bruker) using powdered samples. Asymmetric-flow field-flow fractionation (AF4) (Postnova Analytics) was used to determine molar masses[31]. The flow behaviour of **P1** and **P2** (polymer concentrations equal a charge storage capacity of about 10 A h l^{-1}) in aqueous NaCl solution (1 mol l^{-1}) was studied on an MCR301 rotational rheometer (Anton Paar) at 20 °C under continuous shear using a double-gap measuring system (DG26.7).

Representative synthesis procedures. *Poly(2,2,6,6-tetramethylpiperidinyloxy-4-yl methacrylate-co-[2-(methacryloyloxy) ethyl]trimethylammonium chloride), **P1**.* 2,2,6,6-Tetramethylpiperidin-4-yl-methacrylate **1** (50 g, 222 mmol), [2-(methacryloyloxy)ethyl]trimethylammonium chloride **2** (53.3 ml, 222 mmol, 80% solution in water), and 2-mercaptoethanol (2.5 ml, 35 mmol) were dissolved in dilute hydrochloric acid (295 ml, 0.75 mol l^{-1}). After flushing the solution with argon for 60 min, 4,4′-azobis(4-cyanovaleric acid) (ABCVA) (7.5 g, 22 mmol) was added. The reaction mixture was stirred at 75 °C for 6 h. Subsequently, the solution was cooled to room temperature, and hydrogen peroxide (34 ml, 333 mmol, 30% solution in water), sodium tungstate (0.85 g, 2.9 mmol), EDTA (0.28 g, 1 mmol), and aqueous sodium hydroxide solution (110 ml, 10 wt%) were added. The solution was stirred for 48 h; additional hydrogen peroxide (34 ml, 333 mmol) was added after 24 h. Afterwards, the solution was filtered, dialysed against water (MWCO = 1,000 g mol^{-1}) and lyophilized to yield an orange powder (94 g).

The properties of **P1** are as follows. M_n = 20,200 g mol^{-1}; M_w = 33,700 g mol^{-1}, dispersity $Đ$ = 1.7 (determined by AF4); ^{1}H NMR (D_2O, 300 MHz) δ (in p.p.m.), 4.65 (s, br, 1H), 4.17 (s, br, 2H), 3.44 (s, br, 2H), 2.90 (s, br, 9H), and 2.15–0.25 (m, 26H), (radical quenched by phenylhydrazine); ATR-IR (powder; in cm^{-1}), 3,400 (vb), 2,974 (w), 2,945 (w), 1,721 (s), 1,475 (w), 1,388 (w), 1,236 (w), 1,145 (m), and 952 (w); g = 2.0074 (determined by EPR); capacity, 39 mA h g^{-1} (determined by titration) and 37 mA h g^{-1} (determined by EPR); mean hydrodynamic radius, 2 nm (determined by dynamic light scattering).

*Poly(N-4-vinylbenzyl-N′-methyl-4,4′-bipyridinium dichloride-co-4-vinylbenzyl trimethylammonium chloride), **P2**.* 4-Vinylbenzyl chloride **3** (94.5 g, 620 mmol), 4-vinylbenzyl trimethylammonium chloride **4** (20.2 g, mmol), and 2,2′-azobis(2-methylpropionitrile) (3.5 g, 21 mmol) were dissolved in dimethyl sulfoxide (800 ml). After flushing the solution with argon for 60 min, the reaction mixture was stirred at 75 °C for 6 h. Subsequently, *N*-methyl-4,4′-bipyridinium iodide (195 g, 620 mmol) was added, and the solution was stirred at 80 °C for 48 h and dialysed against water (MWCO = 10,000 g mol^{-1}). Ion exchange from iodide to chloride was performed by Dowex Marathon A exchange resin. The obtained solution was lyophilized to yield an ochre powder (172 g).

The properties of **P2** are as follows. M_n = 30,900 g mol^{-1}; M_w = 73,400 g mol^{-1}, $Đ$ = 2.4 (determined by AF4); ^{1}H NMR (D_2O, 300 MHz) δ (in p.p.m.), 8.95 (m, 4H), 8.43 (s, br, 4H), 7.75–6.20 (m, 4H + 4H), 5.78 (s, br, 2H), 4.40 (s, br, 3H), 2.90 (s, br, 9H), and 2.25–0.20 (m, 3H + 3H); ATR-IR (powder; in cm^{-1}), 3,360 (b), 3,005 (m), 2,926 (m), 1,635 (s), 1,558 (m), 1,506 (m), 1,446 (m), 1,352 (w), 1,217 (w), 823 (m), 796 (m), and 669 (w); capacity, 51 mA h g^{-1} (determined by charging/discharging test); mean hydrodynamic radius, 2 nm (determined by dynamic light scattering).

Electrochemical investigations. *Cyclic voltammetry and RDE voltammetry.* These were conducted on a VersaSTAT potentiostat/galvanostat (Princeton Applied Research) using a standard three-electrode set-up with a glassy-carbon-disk working electrode (5 mm diameter), a Ag/AgCl/water reference electrode, and a graphite rod counter electrode. For RDE voltammetry, the rotation speed was controlled externally by a Model 636A ring-disk electrode system (Princeton Applied Research).

For **P1**, analysis of the RDE voltammograms via a Levich plot (limiting current i_{lim} versus $\omega^{1/2}$, where ω is the rotation speed) yields the corresponding diffusion coefficient D using the Levich equation $i_{lim} = 0.62nFAD^{2/3}\omega^{1/2}\nu^{-1/6}c_0$, where $n = 1$ is the number of transferred electrons per redox reaction, F = 96,485 C mol^{-1} is Faraday's constant, A = 0.20 cm^2 is the area of the electrode surface, $\nu = 1.01 \times 10^{-6}$ m^2 s^{-1} is the kinematic viscosity of the aqueous sodium chloride solution (0.1 mol l^{-1}), and c_0 is the bulk concentration of the redox-active repeating unit of the polymer. Application of the Koutecký–Levich equation $1/i = 1/i_k + 1/i_{lim}$ yields the mass-transfer-independent kinetic current i_k, which is subsequently fitted by the Butler–Volmer equation via a Tafel plot ($\log(|i_k|)$ versus overpotential). This fitting allows us to determine i_0 ($\log[i_k(0)] = \log|i_0|$), and consequently k^0 (via $i_0 = FAk^0c_0$) and α (via the slope of the Tafel plot: $-\alpha F/(2.3RT)$ for negative slopes or $(1 - \alpha)F/(2.3RT)$ for positive slopes, where R is the universal gas constant and T is the absolute temperature).

For **P2**, a two-step process is assumed for the first reduction process on the basis of the observed voltammetric behaviour. Furthermore, the analysis of RDE experiments is restricted to high rotation rates, because a layer of an intermediate product is observed at low rotation rates. For analysis, the respective EE mechanism (which includes two subsequent electrochemical reactions) following ref. 32 is applied. Accordingly, for small negative overpotentials, we apply the Levich equation and determine D as described above. We analyse the current for small negative overpotentials η_1 using $1/i = 1/(2i_{lim}) + \exp[\eta_1 F/(RT)]/(2FAk^0c_0)$, which yields $1/i_k(\eta_1) = \exp[\eta_1 F/(RT)]/(2FAk^0c_0)$ as $\omega^{-1/2} \to 0$ and $\log|1/i_k(\eta_1 = 0)| = -\log|2FAk^0c_0|$; we use this expression to determine k^0. This analysis does not allow us to determine α.

Spectroelectrochemical experiments. These were carried out in a quartz cuvette (optical path length of 1 mm) containing 0.1 mol l^{-1} NaCl in water solution, a platinum-grid working electrode, a platinum-wire auxiliary electrode and a Ag/AgCl/water reference electrode. The potential was controlled using an Autolab PGSTAT30 potentiostat (Metrohm). The redox process was monitored by ultraviolet−visual spectroscopy using a Lambda 750 UV-vis spectrophotometer (PerkinElmer) and considered complete when there was no further spectral change.

Charging/discharging tests. These were carried out at 25 °C using a potentiostat (VMP3, Biologic) and an RFB test cell (JenaBatteries GmbH; see Extended Data Fig. 3): poly(tetrafluoroethylene) (PTFE) frame, ethylene propylene diene monomer (EPDM) rubber seals, graphite (cathode) and Nickel (anode) current collectors, graphite felt electrodes (2.25 × 2.25 × 0.4 cm^3, GFA6, SGL), and an active area of 5 cm^2. For static experiments, the flow-cell set-up was filled with 3 ml electrolyte solution (**P1** at 2 A h l^{-1} and **P2** at 4 A h l^{-1} in 2 mol l^{-1} aqueous NaCl solution) using a syringe and sealed. The effective electrode volume of this unpumped cell is about 2 ml. For dynamic experiments, polymer solutions with a charge storage capacity of 10 A h l^{-1} in aqueous NaCl solution (2 mol l^{-1}) were prepared. 10 ml **P1** solution and 15 ml **P2** solution were transported through the cell by a peristaltic pump at a flow rate of 20 ml min^{-1} (Hei-FLOW Advantage, Heidolph). The battery was charged/discharged under a constant-current regimen. All electrolyte solutions were kept under an argon atmosphere. The electric performance of the polymer-based RFB cell was studied in a pumped cell. For studying the influence of the discharging current density (20–100 mA cm^{-2}), the charging current density was kept constant at 40 mA cm^{-2}; an inverse experiment was conducted for studying the influence of the charging current density. A long-term cycling test was performed by repeatedly charging and discharging a static cell at 20 mA cm^{-2}. The state-of-charge curve was acquired by incremented charge/discharge at 5 mA cm^{-2} for a time of 10 s. A lower cut-off potential of 0.8 V and an upper cut-off potential of 1.35 V was used for all experiments. For all experiments a cellulose-based dialysis membrane with an MWCO of 6,000–8,000 g mol^{-1}, a thickness of 70 µm and a pore size <1 nm (Spectra/Por 1, Spectrum Laboratories) was used. The membrane was pretreated with deionized water before use.

Membrane characterization. *Salt permeability.* We determined the salt permeability (P_s) using a homemade set-up consisting of two chambers (A and B) separated by a dialysis membrane (regenerated cellulose; MWCO of 6,000–8,000 g mol^{-1}; thickness of 70 µm; Spectra/Por 1, Spectrum Laboratories)[33]. Chamber A was filled with sodium chloride feed solution (1.0 mol l^{-1}) and chamber B with deionized water (25 °C). We determined the change of the salt concentration in chamber B via conductivity measurements. Subsequently, we calculated the salt permeability from the change of the salt concentration over time from two averaged runs using $P_s = \frac{V_B}{c_A A}\frac{dc_B}{dt} = \frac{D_s}{L}$, where $V_{A,B}$ are the volumes of the solutions in chambers A and B, A is the membrane area, L the membrane thickness, $c_{A,B}$ are the salt concentrations in chambers A and B, t is time, and D_s is the diffusion coefficient of the salt.

We determined the area resistance R using a 5-cm^2 test cell. The data represent an average of three measurements. We measured the electrolyte resistance (aqueous NaCl, 1 mol l^{-1}) of the cell with (R_1) and without (R_2) a membrane using electrochemical impedance spectroscopy; R was calculated as $R = (R_1 - R_2)A$.

Retention of redox-active polymers. We studied the retention of the redox-active polymers in an RFB test cell (JenaBatteries GmbH; see Extended Data Fig. 3) at room temperature by pumping a 10 ml feed solution of **P1** (60 mg ml^{-1} in 2.0 mol l^{-1} $NaCl_{aq}$) or **P2** (40 mg ml^{-1} in 2.0 mol l^{-1} $NaCl_{aq}$) through one cell compartment and 10 ml of pristine sodium chloride solution (2.0 mol l^{-1}) through the second cell compartment. Both compartments were separated by a dialysis membrane (Spectra/Por 1, Spectrum Laboratories). We analysed the polymer

concentration in the second cell compartment by ultraviolet–visual spectroscopy using a Lambda 750 UV–vis spectrophotometer (PerkinElmer).

Maximum polymer permeability. We calculated the maximum polymer permeability ($P_{polymer,max}$) in accordance with P_s. In contrast to permeability data for molecules with a defined molar mass (for example, metal ions, or 'small' organic molecules), the polymer permeability cannot be expressed with one absolute value because polymers have a molar mass distribution. P_s decreases substantially as the molar mass of a single molecule increases. Therefore, only $P_{polymer,max}$ for the smallest polymer molecules of a given distribution can be determined. Larger polymer molecules will pass the membrane at a much lower rate; that is, they have a much smaller polymer permeability.

Membrane selectivity. Membrane selectivity (S) describes the ratio of the rates of ion transfer of the electrolyte salt (H^+ for a vanadium-based RFB or NaCl for a polymer-based RFB) and the redox-active species (vanadium salts for a vanadium-based RFB or **P1** and **P2** for a polymer-based RFB), according to $S = P_{s,NaCl}/P_{polymer,max}$. In the context of determining a maximum permeability for polymers, membrane selectivity is considered a minimum value (S_{min}) for polymers.

Cost calculations. The data on Nafion membrane prices indicated a current cost of US$500–1,000 m^{-2} and a projected long-term cost of US$100–200 m^{-2} (refs 34–36). Dialysis and nanofiltration membranes are currently available at US$20–100 m^{-2} on an industrial scale from manufactures such as Spectrum Laboratories, Pall, Microdyn-Nadir, Sartorius and Merck Millipore (for example http://www.spectrumlabs.com/dialysis/RCtubing.html). Future expansion in production capabilities and learning-curve effects will further reduce the price (probably to <US$10 m^{-2}).

Similar to the membrane material, the electrolyte contributes substantially to the cost of RFBs. Polymers are ubiquitous and can be prepared at very low costs on a megatonne scale. The prices vary between about US$0.6 kg^{-1} for commodity plastics (for example, polyvinyl chloride or polystyrene), US$1.2 kg^{-1} for engineering plastics (for example, nylon or polymethyl methacrylate), and US$1.5–3 kg^{-1} for high-performance plastics (for example, polyetheretherketone or polyethylenimine); http://plasticker.de, accessed July 2015. Redox-active polymers can be prepared at a price similar to high-performance plastics on a kilotonne scale. Because large photovoltaic and wind farms require megawatt batteries, which use tonnes of active material, production numbers will soon rise to a level that allows for economical polymer synthesis. Upon further increases in the energy density of the polymers and industrial up-scaling of the production, we anticipate a competitive price for polymer-based RFBs, which benefit from the less-corrosive electrolyte.

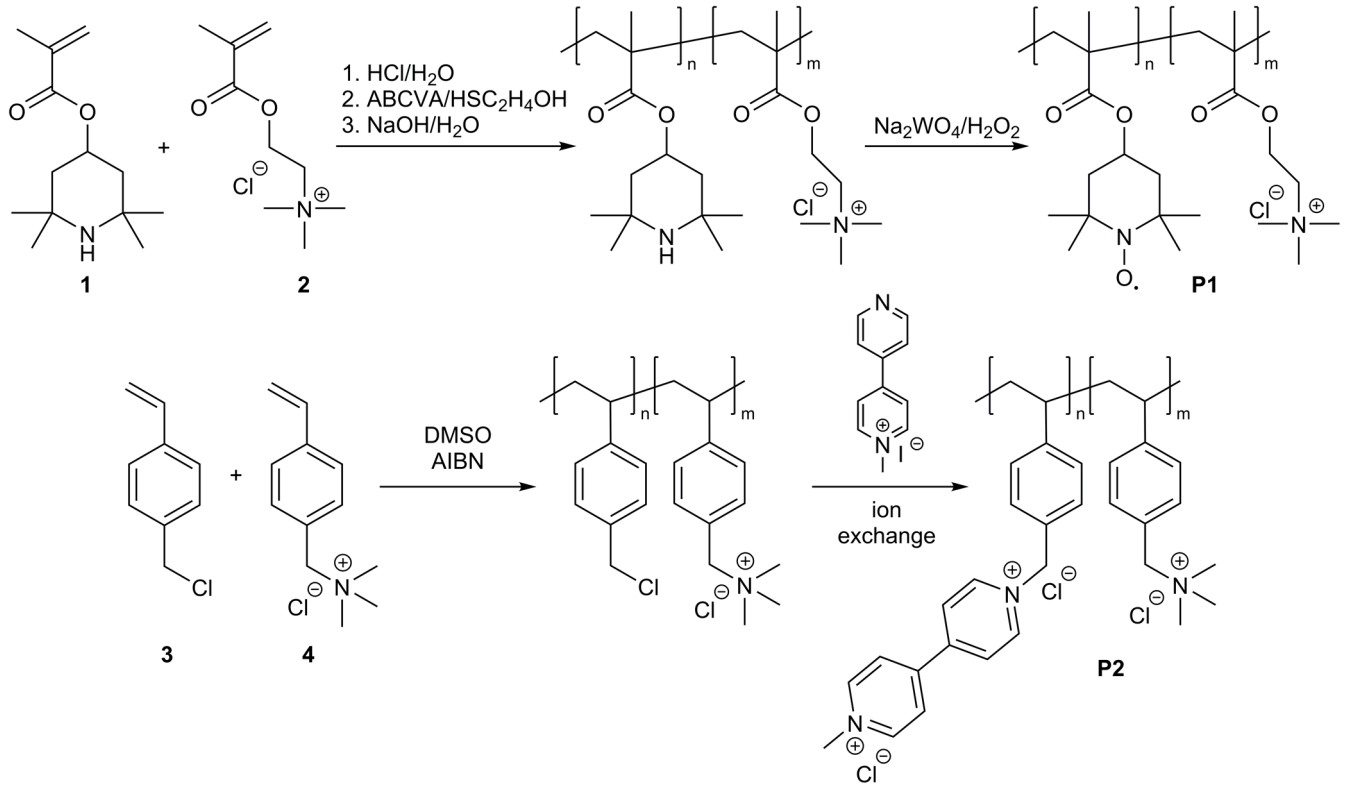
Dynamic light scattering. Dynamic light-scattering measurements were performed on an ALV CGS-3 (Malvern) equipped with a He–Ne laser (633 nm) at polymer concentrations of 5 mg ml^{-1} at 25 °C. We analysed the experimental autocorrelation functions using the CONTIN algorithm. We calculated the apparent hydrodynamic radii using the Stokes–Einstein equation.

Toxicity tests. Cytotoxicity studies were performed with the mouse fibroblast cell line L929 (CCL-1, ATCC), as recommended by ISO10993-5. The cells were routinely cultured in DMEM, supplemented with 10% fetal calf serum, 100 U ml^{-1} penicillin, and 100 μg ml^{-1} streptomycin (all components from Biochrom), at 37 °C in a humidified 5% (v/v) CO_2 atmosphere.

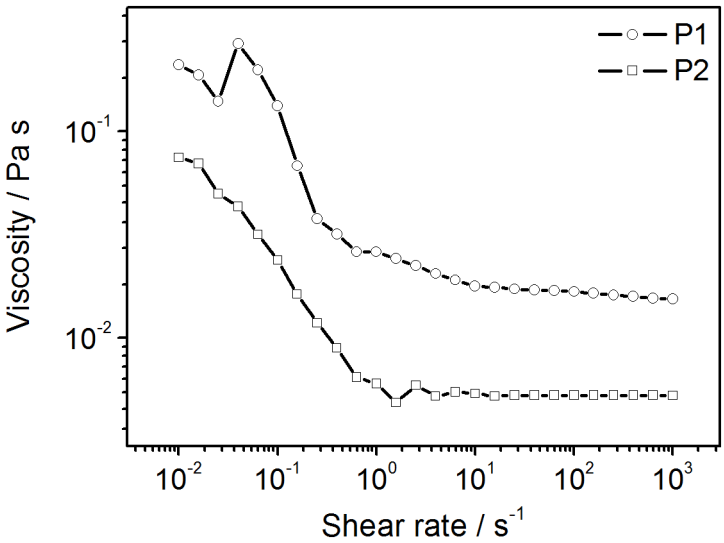
Cells were seeded at 10^4 cells per well in a 96-well plate and incubated for 24 h; no cells were seeded in the outer wells. Afterwards, the testing substances were added to the cells at concentrations indicated in Extended Data Fig. 7 (0.25 μg ml^{-1} to 1 mg ml^{-1}) and the plates were incubated for a further 24 h. Polymers **P1** and **P2** were applied in a diluted aqueous sodium chloride solution as used for the charging/discharging tests, and the vanadium salts were used in a diluted sulphuric acid solution with concentration ratios commonly used in vanadium-based RFBs (1.5 mol l^{-1} vanadium ion, 3.5 mol l^{-1} sulfuric acid). Control cells were incubated with fresh culture medium. Subsequently, the medium was replaced by a mixture of fresh culture medium and Alamar-Blue solution (Life Technologies), prepared according to the manufacturer's instructions. After a further incubation period of 4 h at 37 °C, the fluorescence was measured at excitation/emission wavelengths of 570 nm/610 nm, with untreated cells on the same well plate serving as negative controls. The negative control was standardized as 0% of metabolism inhibition and referred as 100% viability. Cell viability below 70% was considered indicative of cytotoxicity. Data are expressed as mean ± s.d. of three determinations.

The membrane damaging properties of polymers were quantified by analysing the haemoglobin release from erythrocytes by a haemolysis assay. Blood from sheep, collected in heparinized-tubes (Institut für Versuchstierkunde und Tierschutz, Friedrich Schiller University Jena), was centrifuged at 4,500*g* for 5 min, and the pellet was washed three times with cold 1.5 mmol l^{-1} phosphate buffered saline (PBS; pH 7.4). After dilution with PBS in a ratio of 1:7, aliquots of erythrocyte suspension were mixed 1:1 with the polymer solution and incubated in a water bath at 37 °C for 60 min. After centrifugation at 2,400*g* for 5 min, the haemoglobin release into the supernatant was determined spectrophotometrically using a microplate reader (TECAN Infinite M200 PRO plate reader) at a wavelength of 544 nm. Complete haemolysis (100%) was achieved using 1% Triton X-100 serving as the positive control; PBS served as negative control (0%). A haemolysis rate less than 2% was taken as non-haemolytic. Experiments were run in triplicates and were performed with blood from three different animals.

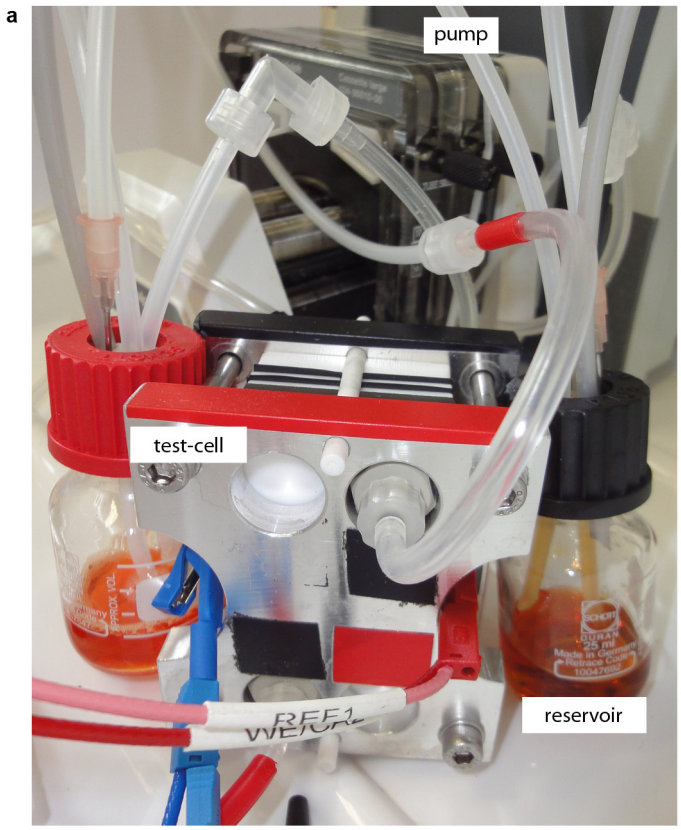
30. Park, Y. S., Lee, E. J., Chun, Y. S., Yoon, Y. D. & Yoon, K. B. Long-lived charge-separation by retarding reverse flow of charge-balancing cation and zeolite-encapsulated $Ru(bpy)_3^{2+}$ as photosensitized electron pump from zeolite framework to externally placed viologen. *J. Am. Chem. Soc.* **124,** 7123–7135 (2002).
31. Wagner, M., Pietsch, C., Tauhardt, L., Schallon, A. & Schubert, U. S. Characterization of cationic polymers by asymmetric flow field-flow fractionation and multi-angle light scattering—a comparison with traditional techniques. *J. Chromatogr. A* **1325,** 195–203 (2014).
32. Treimer, S., Tang, A. & Johnson, D. C. A consideration of the application of Koutecký–Levich plots in the diagnoses of charge-transfer mechanisms at rotated disk electrodes. *Electroanalysis* **14,** 165–171 (2002).
33. Cañas, A., Ariza, M. J. & Benavente, J. A comparison of electrochemical and electrokinetic parameters determined for cellophane membranes in contact with NaCl and $NaNO_3$ solutions. *J. Colloid Interface Sci.* **246,** 150–156 (2002).
34. Zhang, M., Moore, M., Watson, J. S., Zawodzinski, T. A. & Counce, R. M. Capital cost sensitivity analysis of an all-vanadium redox-flow battery. *J. Electrochem. Soc.* **159,** A1183–A1188 (2012).
35. Viswanathan, V. *et al.* Cost and performance model for redox flow batteries. *J. Power Sources* **247,** 1040–1051 (2014).
36. Minke, C. & Turek, T. Economics of vanadium redox flow battery membranes. *J. Power Sources* **286,** 247–257 (2015).
37. Oh, S. H. *et al.* A metal-free and all-organic redox flow battery with polythiophene as the electroactive species. *J. Mater. Chem. A* **2,** 19994–19998 (2014).

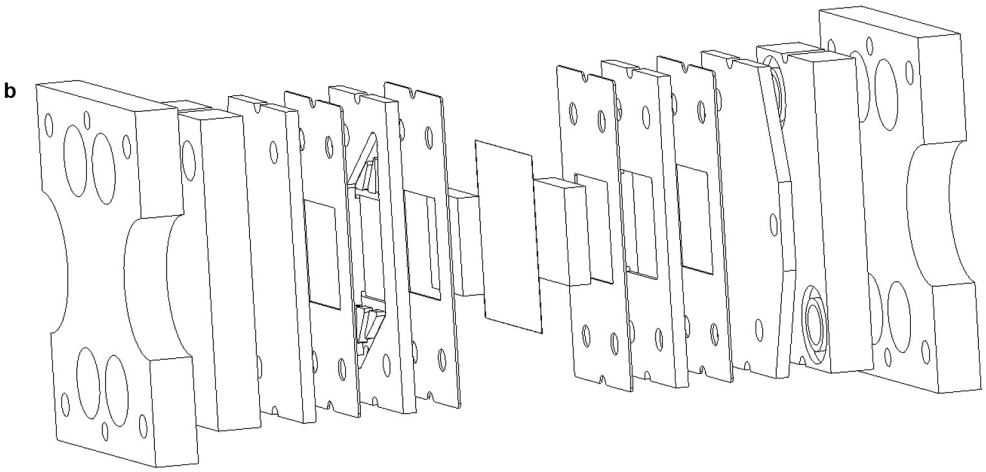


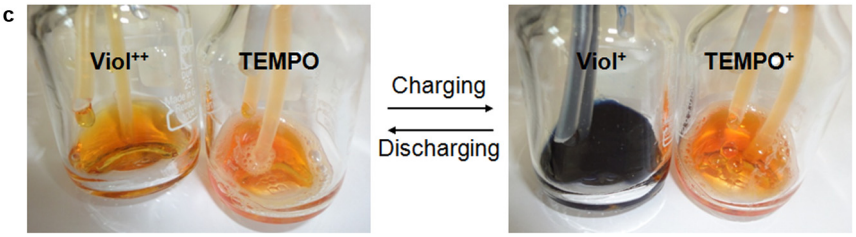
Extended Data Figure 1 | Schematic representation of the synthesis of redox-active polymers. The cathode material **P1** and anode material **P2** were prepared by free radical polymerization and subsequent polymer-analogous oxidation and functionalization, respectively. ABCVA, 4,4′-azobis(4-cyanovaleric acid); AIBN, azobisisobutyronitrile; DMSO, dimethyl sulfoxide.



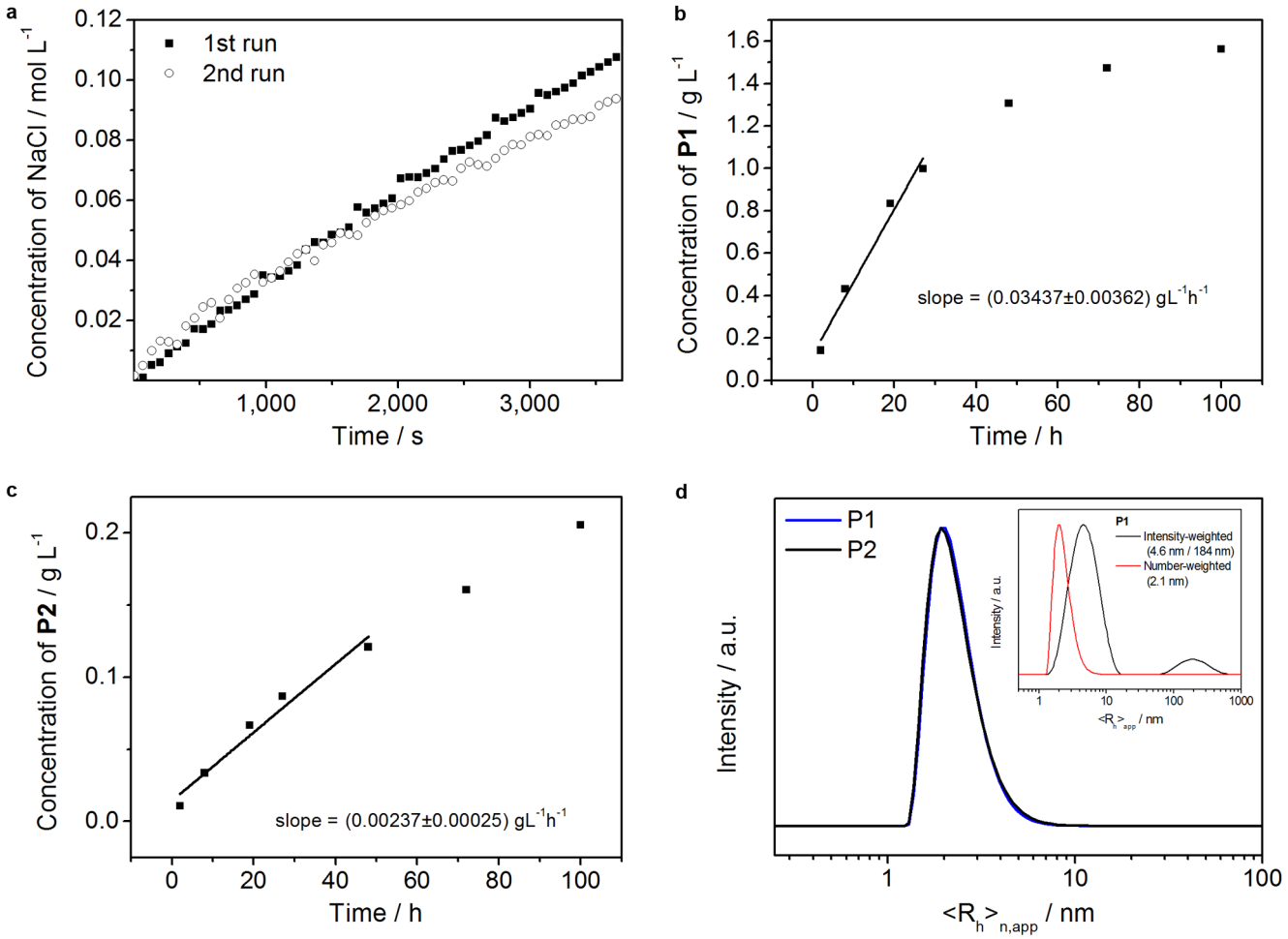
Extended Data Figure 2 | Rheogram of redox-active polymers. The flow behaviour of aqueous solutions of **P1** and **P2** was studied at 20 °C under continuous shear in sodium chloride solution (1 mol l^{-1}) using a double-gap measuring system in a rotational rheometer. The concentrations of the polymers correspond to a charge-storage capacity of about 10 A h l^{-1}.



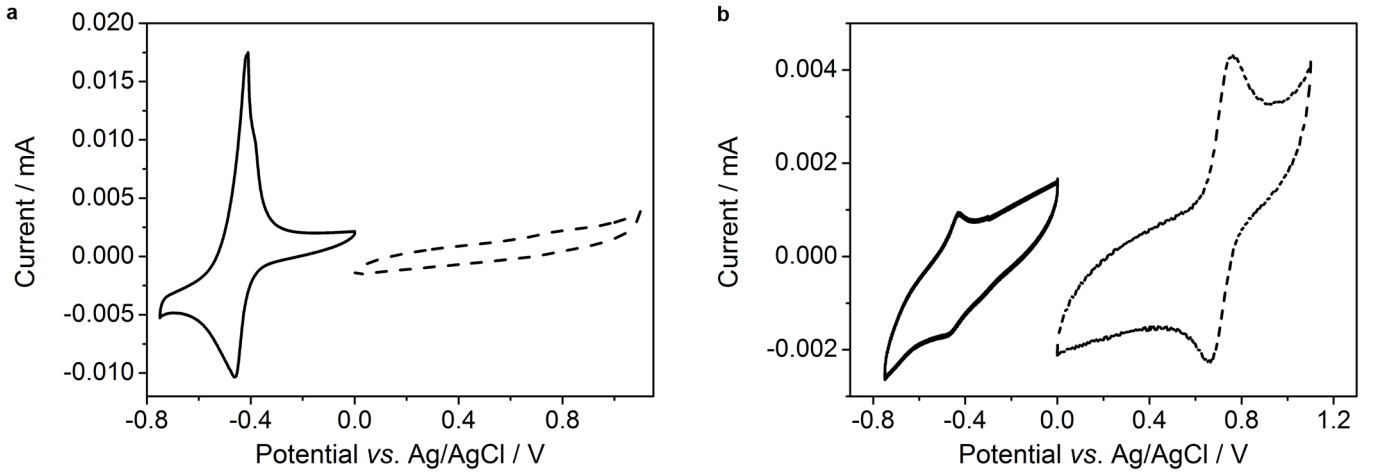




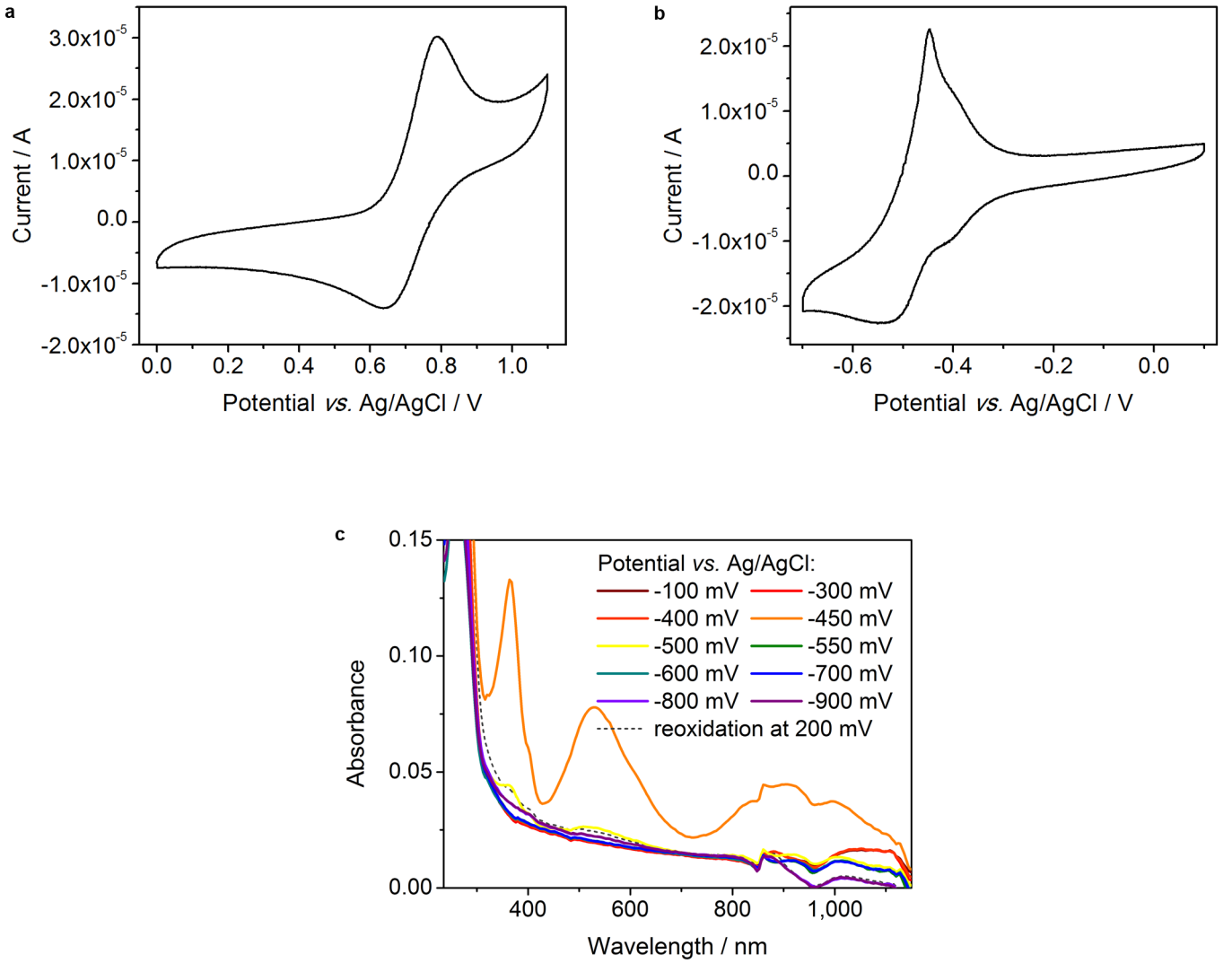
Extended Data Figure 3 | Test set-up for a polymer-based RFB. **a**, Photograph of a laboratory set-up (5-cm^2 test cell, peristaltic pump, and electrolyte reservoirs) used for charging/discharging experiments. **b**, Exploded-view drawing of the 5-cm^2 test cell. **c**, Colour change of the polymer solutions upon charging and discharging.



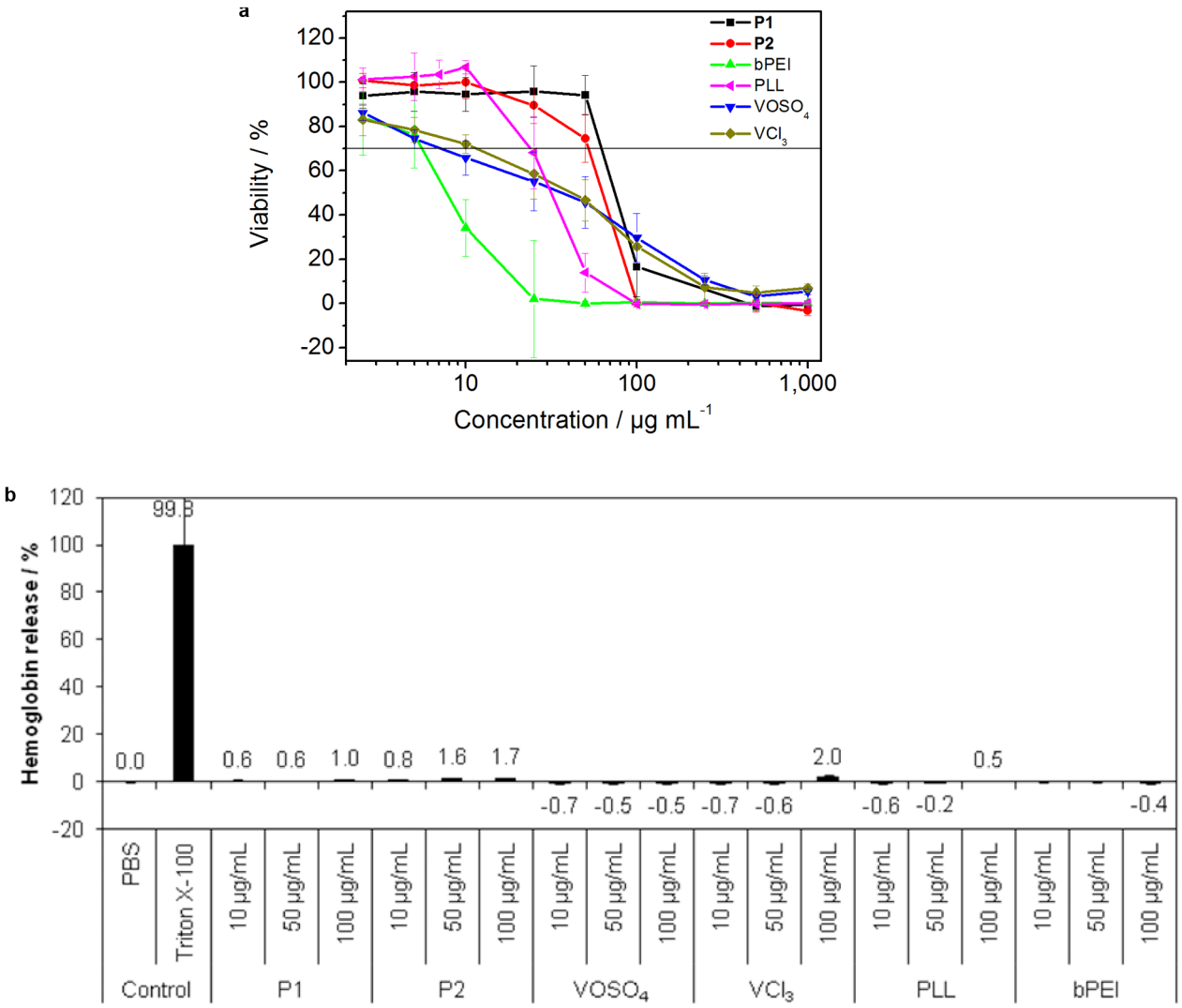
Extended Data Figure 4 | Crossover studies on dialysis membrane. **a**, Time-dependent NaCl concentration of a chamber filled with deionized water that is separated from a NaCl feed solution (1 mol l^{-1}) by a dialysis membrane. Salt permeability was determined to be $P_s = (9.3 \pm 0.1) \times 10^{-5}$ cm s^{-1}. **b**, Time-dependent **P1** concentration of an RFB test cell compartment filled with NaCl solution (2 mol l^{-1}) that is separated from a **P1** feed solution (60 mg ml^{-1}) by a membrane. We determined the maximum polymer permeability from the linear part of the diffusion graph to be $P_{polymer,max} = (3.2 \pm 0.3) \times 10^{-7}$ cm s^{-1} (diffusion coefficient $D_{polymer,max} = (1.3 \pm 0.1) \times 10^{-7}$ cm^2 min^{-1}). The minimum membrane selectivity $S_{min} = 290$. **c**, Time-dependent change in **P2** concentration of an RFB test cell compartment filled with NaCl solution (2 mol l^{-1}) that is separated from a **P2** feed solution (40 mg ml^{-1}) by a dialysis membrane. $P_{polymer,max} = (3.3 \pm 0.4) \times 10^{-8}$ cm s^{-1}; $D_{polymer,max} = (1.4 \pm 0.2) \times 10^{-8}$ cm^2 min^{-1}; $S_{min} = 2{,}830$. All crossover experiments were conducted with a cellulose-based dialysis membrane (MWCO = 6,000 g mol^{-1}) at 25 °C. In **b** and **c**, the slopes of the fit lines correspond to $\frac{dc_B}{dt}$ in $P_s = \frac{V_B}{c_A A}\frac{dc_B}{dt} = \frac{D_s}{L}$ (see Methods). **d**, The number-weighted distributions of hydrodynamic radii ($\langle R_h \rangle_{n,app}$) of **P1** and **P2** determined by dynamic light scattering reveal mean radii of approximately 2 nm (5 g l^{-1} in 0.1 mol l^{-1} NaCl solution). Inset, a comparison of the intensity- and number-weighted distributions of **P1** shows the presence of aggregates (4.6 nm and 184 nm).



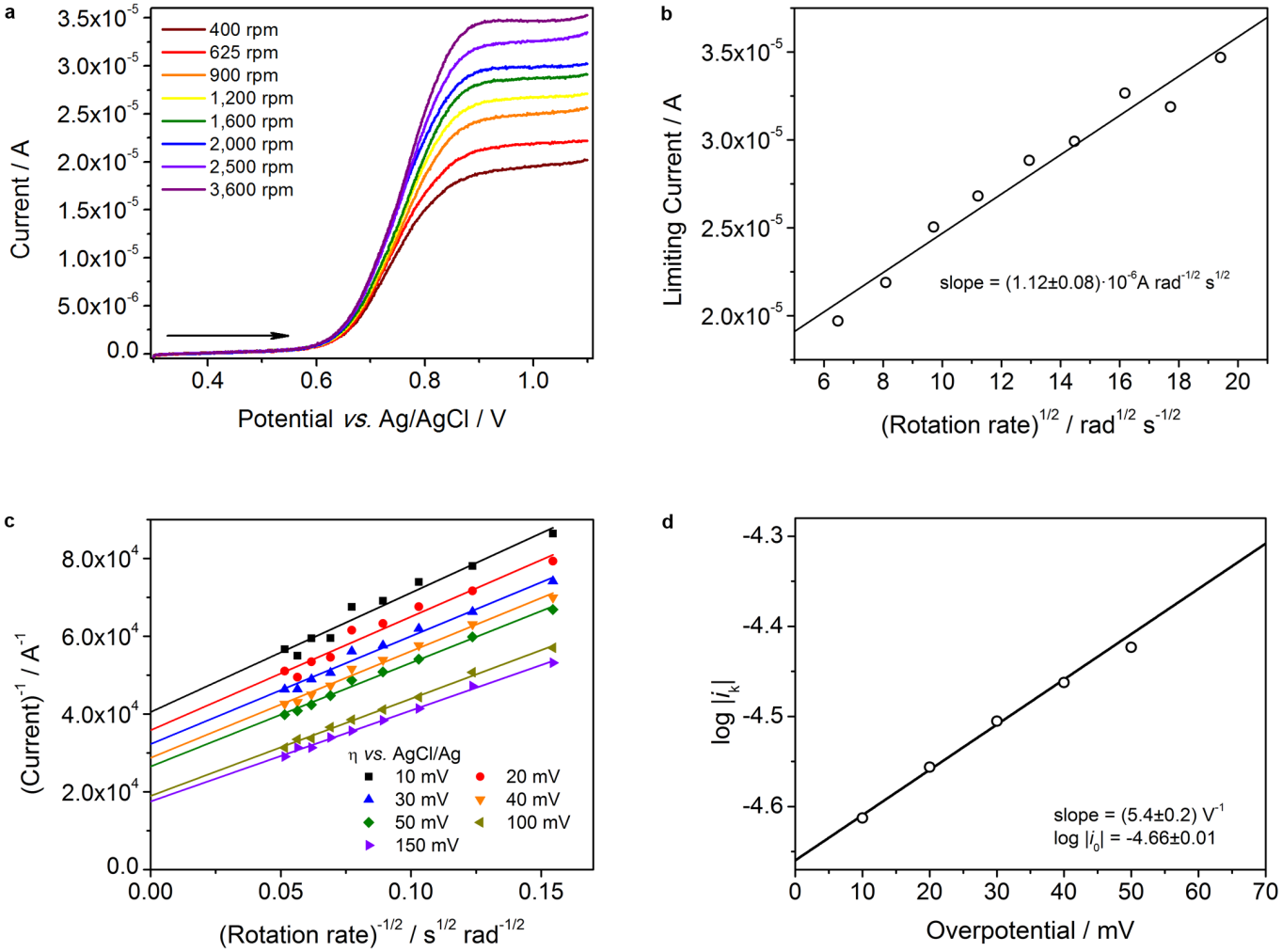
Extended Data Figure 5 | Cyclic voltammogram of electrolytes after 10,000 cycles. **a**, **b**, The cyclic voltammogram (in water with 0.1 mol l^{-1} NaCl; scan rate of 200 mV s^{-1}) of samples taken from the anolyte (**a**) and catholyte (**b**) after repeated charging/discharging; solid lines and dashed lines correspond to the reductive and oxidative range, respectively, of the cyclic voltammogram.



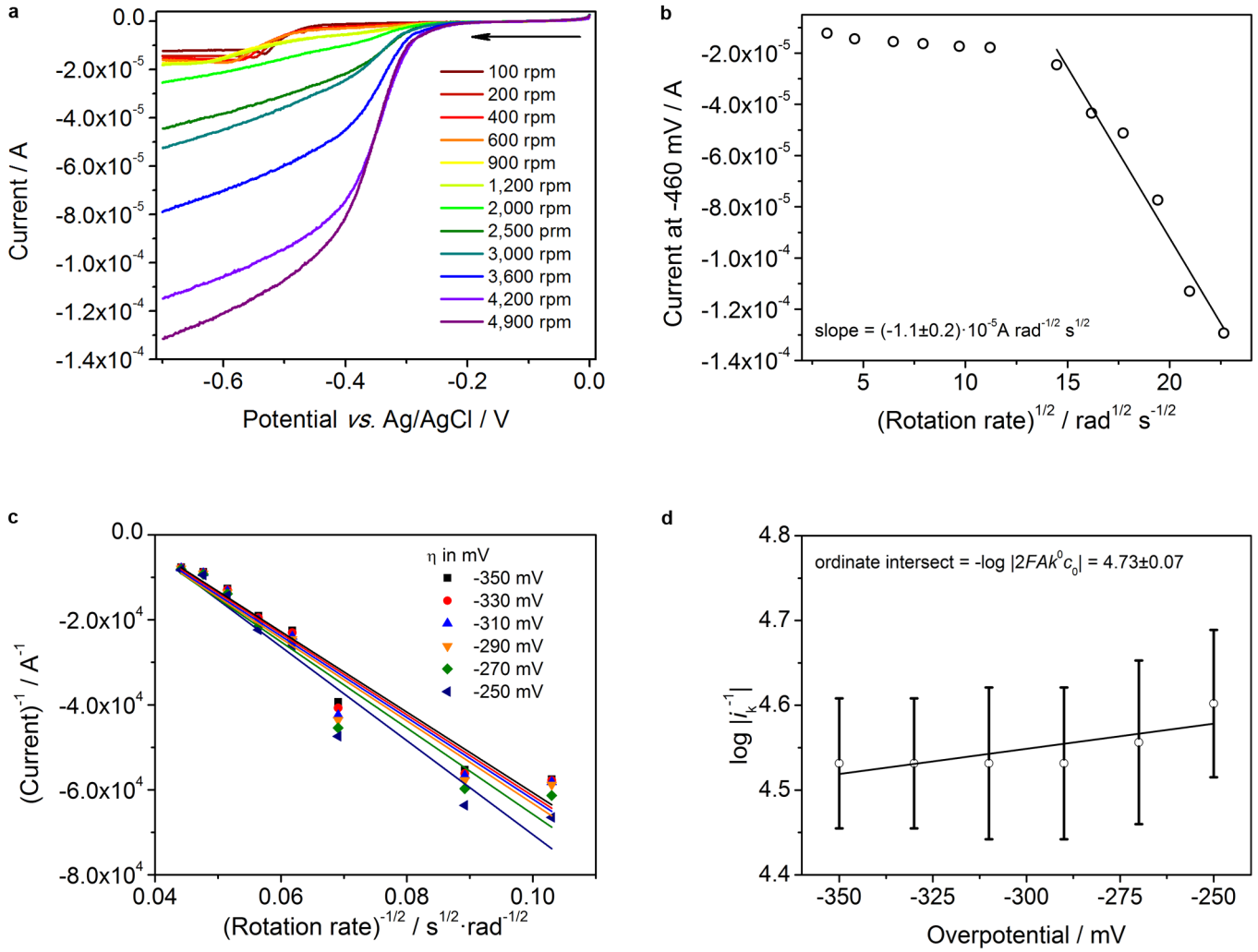
Extended Data Figure 6 | Electrochemical analysis of P1 and P2. **a**, Cyclic voltammogram of the oxidation process of **P1** (2.5×10^{-3} mol l^{-1} in water with 0.1 mol l^{-1} NaCl; scan rate of 200 mV s^{-1}). **b**, Cyclic voltammogram of the first reduction process of **P2** (5.2×10^{-3} mol l^{-1} in water with 0.1 mol l^{-1} NaCl; scan rate of 200 mV s^{-1}). **c**, Ultraviolet–visual spectroelectrochemistry of **P2** at different applied potentials (as indicated) during consecutive reduction and after subsequent re-oxidation (10^{-4} mol l^{-1} in water with 0.1 mol l^{-1} NaCl). The single-reduced species (Viol$^{+\bullet}$) is visible at −450 mV (orange line) with the formation of three distinct bands at 365 nm, 530 nm, and 900 nm, which disappear upon further reduction towards the double-reduced species (Viol^{0}). The shapes and positions of the emerging bands strongly suggest the formation of radical cation dimers[27]. Re-oxidation at 200 mV (dotted line) restores the initial spectrum.



Extended Data Figure 7 | Toxicity tests of the redox-active polymers. **a**, The viability of L929 mouse cells was tested in the presence of redox-active compounds according to ISO10993-5. Cell viability below 70% is considered indicative of cytotoxicity. The negative control was standardized as 0% of metabolism inhibition and referred as 100% viability. Two vanadium salts in redox states commonly found in vanadium-based RFBs and two widely used cationic polymers were used as a reference. Poly(L-lysine), PLL, is a commercial food preservative and branched poly(ethylene imine), bPEI, is used in the paper-making industry and as a flocculating agent. Although **P1** and **P2** show cytotoxic effects at concentrations $>50\ \mu g\ ml^{-1}$—with **P1** being less toxic than **P2**—the vanadium salts and the cationic polymers reveal cytotoxic effects at lower concentrations ($VOSO_4 > 5\ \mu g\ ml^{-1}$, $VCl_3 > 10\ \mu g\ ml^{-1}$, bPEI $> 5\ \mu g\ ml^{-1}$, PLL $> 25\ \mu g\ ml^{-1}$). Data are expressed as mean values and error bars represent the standard deviation of three determinations. **b**, We quantified the cell-membrane damaging properties of the polymers by analysing the haemoglobin release from erythrocytes (indicated by the numbers associated with each bar). Data are expressed as mean values and error bars represent the standard deviation of triplicates of three different blood samples per concentration. Because a haemoglobin release (haemolysis rate) less than 2% is considered non-haemolytic, **P1** and **P2** as well as the reference compounds show no membrane damaging behaviour. Hence, the cell toxic effects do not originate from damage to the cell membrane, but from reactions within the cell. Because the cell uptake via diffusive processes of polymers is hindered in comparison to 'small' inorganic ions, **P1** and **P2** possess lower cytotoxicity. These tests provide some insight into the toxicity, but long-term ecotoxicity tests and animal testing are required to fully evaluate the impact of the redox-active polymers on wildlife and plants.

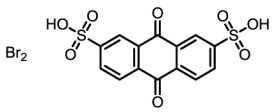
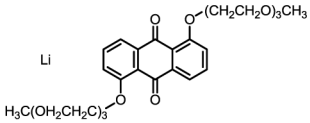
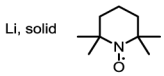
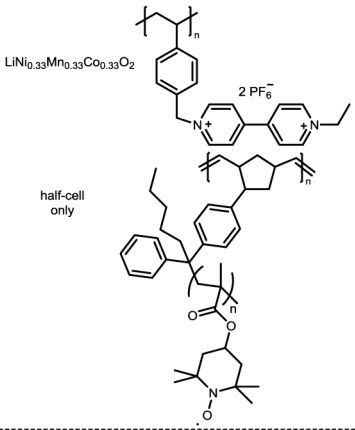
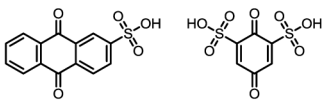
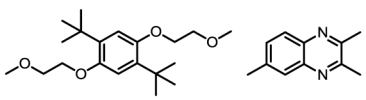
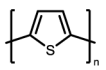
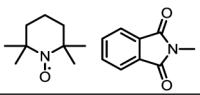


Extended Data Figure 8 | RDE measurements of P1 and analysis. **a**, Voltammograms of **P1** (2.5×10^{-3} mol l^{-1} in water with 0.1 mol l^{-1} NaCl; scan rate of 5 mV s^{-1}) at different rotation speeds (as indicated) from 400 r.p.m. to 3,600 r.p.m. (The arrow indicates the direction of potential scanning). **b**, Levich plot from the obtained limiting currents; application of Levich equation yields a diffusion coefficient $D = (7.0 \pm 0.5) \times 10^{-8}$ cm^2 s^{-1}. **c**, Koutecký–Levich plot for different overpotentials η (as indicated) yielding the mass-transfer-independent current i_k (as $\omega^{-1/2} \rightarrow 0$, $i_k = i$). **d**, Tafel plot yielding $k^0 = (4.5 \pm 0.1) \times 10^{-4}$ cm s^{-1} and $\alpha = 0.68 \pm 0.03$ (see Methods).



Extended Data Figure 9 | RDE measurements of P2 and analysis. **a**, Voltammograms of **P2** (5.2×10^{-3} mol l^{-1} in water with 0.1 mol l^{-1} NaCl; scan rate 5 mV s^{-1}) at different rotation speeds (as indicated) from 100 r.p.m. to 4,900 r.p.m.; substantial changes of the limiting current, necessary for reasonable analysis, are observed only at rotation speeds >1,200 r.p.m. (The arrow indicates the direction of potential scanning). **b**, Levich plot for currents between the first and second steps of the two-step process; the Levich equation was applied only for high rotation rates, that is, in the region of substantial changes of the limiting current, yielding a diffusion coefficient $D = (7.6 \pm 0.9) \times 10^{-7}$ cm^2 s^{-1}. The fit curve and its slope correspond to the Levich equation (see Methods). **c**, Plot of i^{-1} versus $\omega^{-1/2}$ for high negative overpotentials η (as indicated) yielding $1/i_k$ (as $\omega^{-1/2} \rightarrow 0$, $1/i_k = 1/i$). **d**, Plot of $\log|1/i_k|$ versus η_1 (overpotential with respect to the first step) yielding $-\log|2FAk^0c_0|$ ($\log|1/i_k|$ for $\eta_1 = 0$), which allows us to determine $k^0 = (9 \pm 2) \times 10^{-5}$ cm s^{-1}. The error bars represent error that originates from the linear regression analysis of **c**.

Extended Data Table 1 | (Semi)organic redox-flow batteries

Anode & cathode	Electrolyte	Current density [mA cm^{-2}]	Energy density [Wh L^{-1}]	Cycles	Membrane	Source
	HBr $H_2SO_{4(aq)}$	200–500	16	20	Nafion	[6]
	$LiPF_6$/PC	0.1–10	25	9	PP separator (Celgard)	[17]
	$LiPF_6$/MC	1–10	70–130	100	PE separator	[8]
	$LiBF_4$/ CH_3CN	-	0.3	6	PP/PE separator (Celgard)	[18]
	$(C_4H_9)_4NClO_4$/ MC	-	-	2	PP/PE separator (Celgard)	[20]
	$H_2SO_{4(aq)}$	2–10	5	12	Nafion	[7]
	$LiBF_4$/PC	0.06	2	30	Nafion	[9]
	$(C_2H_4)_4NBF_4$/PC	0.1–5	7	20	Fumasep anion-exchange	[37]
	$NaClO_4$/ CH_3CN	0.35	3.5	20	Nepem cation-exchange	[19]
P1 & **P2**	NaCl/H_2O	40-100	10	10,000	Dialysis membrane	-

(PC = propylene carbonate; MC = mixture of organic carbonates; PP = polypropylene; PE = polyethylene)

Typical performance data of literature examples (refs 6–9, 17–20, 37) of organic RFBs that use conventional membranes and electrolytes.

LETTER

doi:10.1038/nature15738

Spontaneous droplet trampolining on rigid superhydrophobic surfaces

Thomas M. Schutzius[1]*, Stefan Jung[1]*, Tanmoy Maitra[1], Gustav Graeber[1], Moritz Köhme[1] & Dimos Poulikakos[1]

Spontaneous removal of condensed matter from surfaces is exploited in nature and in a broad range of technologies to achieve self-cleaning[1,2], anti-icing[3–6] and condensation control[7,8]. But despite much progress[5–7,9–14], our understanding of the phenomena leading to such behaviour remains incomplete, which makes it challenging to rationally design surfaces that benefit from its manifestation[15–18]. Here we show that water droplets resting on superhydrophobic textured surfaces in a low-pressure environment can self-remove through sudden spontaneous levitation and subsequent trampoline-like bouncing behaviour, in which sequential collisions with the surface accelerate the droplets. These collisions have restitution coefficients (ratios of relative speeds after and before collision) greater than unity[19] despite complete rigidity of the surface, and thus seemingly violate the second law of thermodynamics. However, these restitution coefficients result from an overpressure beneath the droplet produced by fast droplet vaporization while substrate adhesion and surface texture restrict vapour flow. We also show that the high vaporization rates experienced by the droplets and the associated cooling can result in freezing from a supercooled state[20,21] that triggers a sudden increase in vaporization, which in turn boosts the levitation process. This effect can spontaneously remove surface icing by lifting away icy drops the moment they freeze. Although these observations are relevant only to systems in a low-pressure environment, they show how surface texturing can produce droplet–surface interactions that prohibit liquid and freezing water-droplet retention on surfaces.

We examine the motion of a water droplet (typical radius $R_0 \approx 0.1$ cm) initially resting on a superhydrophobic surface in a low-pressure environment via high-speed imaging (Fig. 1a, b). The surface used is a silicon micropillar array, a standard, well-controlled surface platform, treated with a fluorosilane coating; the combination of texture and surface chemistry makes the surface superhydrophobic (Fig. 1a, inset, and Extended Data Table 1; the pillar diameter d, pitch l and height h are 1.4 μm, 6.5 μm and 4.8 μm, respectively). While the droplet is resting on this surface, sudden spontaneous droplet motion and bouncing is observed from side-view imaging (Fig. 1a) once the environmental pressure is reduced—here at a rate of -0.1 bar s^{-1} to approximately 0.01 bar (low vacuum; see Supplementary Video 1). Figure 1b quantifies the motion by plotting the droplet centroid position y against time t. This behaviour is reminiscent of an under-damped, forced, mass–spring–damper system operating at a resonance condition (see Supplementary Video 2, Methods section 'Modelling droplet trampolining' and Extended Data Figs 1 and 2), where mass loss from vaporization does not play a role (Methods section 'Droplet mass loss'). The defining feature of the observed behaviour is that each collision between the droplet and the fully rigid substrate results in momentum gain, as in the case of a bouncing trampolinist (although the trampolinist, in contrast to the droplet, interacts with a fully elastic substrate). Similarities also exist with the Leidenfrost effect[22], in that both involve a vaporizing droplet supported on a vapour cushion with excess pressure caused by a small droplet–substrate gap that restricts the draining-vapour flow; but although for the Leidenfrost effect the thickness of this gap is determined by balancing the droplet weight and pressure forces[23], in our system it is determined by the surface texture, represented by pillar height.

Figure 2b shows plots of y as a function of t and Fig. 2a the associated image sequences for droplets with similar velocities v_1 impacting the same superhydrophobic surface in standard-pressure (approximately 1.0 bar) and low-pressure (approximately 0.01 bar) environments (see also Supplementary Video 3). The variables y and t are non-dimensionalized with respect to the initial droplet radius R_0 and the inertial-capillary timescale $\tau \equiv \sqrt{m/\sigma}$, respectively, where m is the

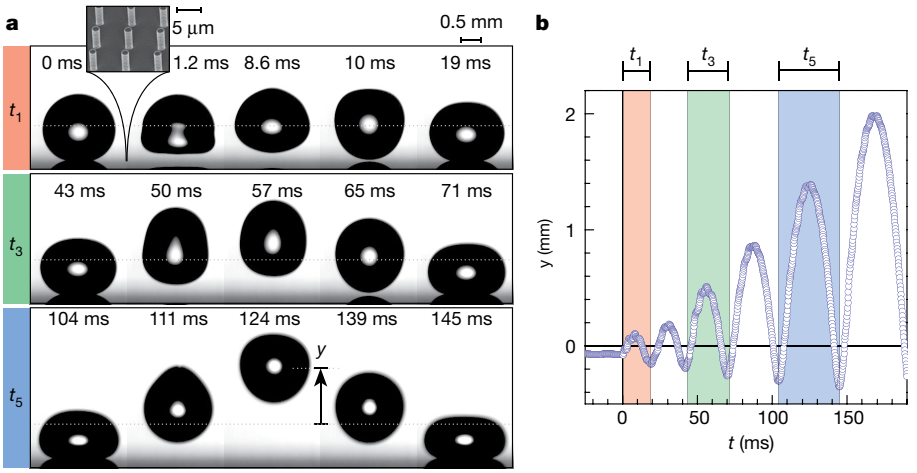


Figure 1 | Droplet trampolining on a rigid surface. a, High-speed image sequence showing a droplet, initially at rest, trampolining (initial droplet radius $R_0 = 0.09$ cm). Inset, micrograph of the silicon superhydrophobic surface showing its pillar texture. **b**, Droplet vertical position y (blue circles, cross-sectional centroid) as a function of time t for the image sequence in **a**. The dotted lines in **a** correspond to $y = 0$ (Extended Data Fig. 1). For more details see Supplementary Video 2.

[1]Laboratory of Thermodynamics in Emerging Technologies, Department of Mechanical and Process Engineering, ETH Zurich, 8092 Zurich, Switzerland.
*These authors contributed equally to this work.

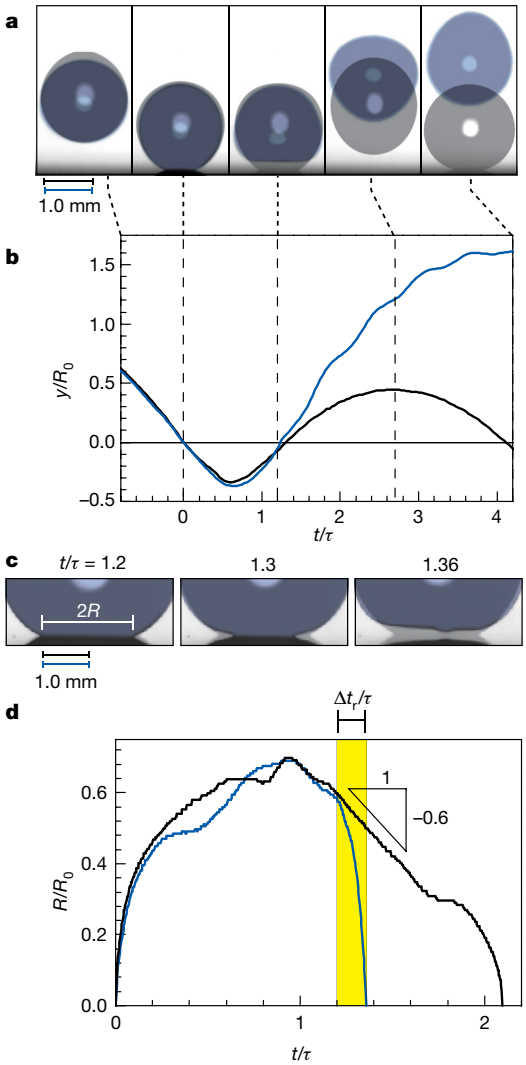


Figure 2 | Vaporization can accelerate droplet recoil. a, Overlaid, semitransparent image sequences of droplets impacting and recoiling from a superhydrophobic surface under environmental conditions at low (blue droplet) and standard (black droplet) pressure. **b**, Plot of y as a function of t (non-dimensionalized with respect to the initial droplet radius R_0 and the inertial-capillary timescale $\tau \equiv \sqrt{m/\sigma}$, respectively) for the image sequences in **a** (blue line, blue droplet; black line, black droplet). **c**, Image sequences similar to those in **a**, but focusing on the contact line region. **d**, Plot of R as a function of t (non-dimensionalized with respect to R_0 and τ, respectively) for the image sequences in **c** (blue line, blue droplet; black line, black droplet). Impact parameters: **a**, **b**, $v_1 = -0.9R_0/\tau$; **c**, **d**, $v_1 = -0.6R_0/\tau$. Surface properties, same as Fig. 1. See Supplementary Videos 3 and 4 for further information.

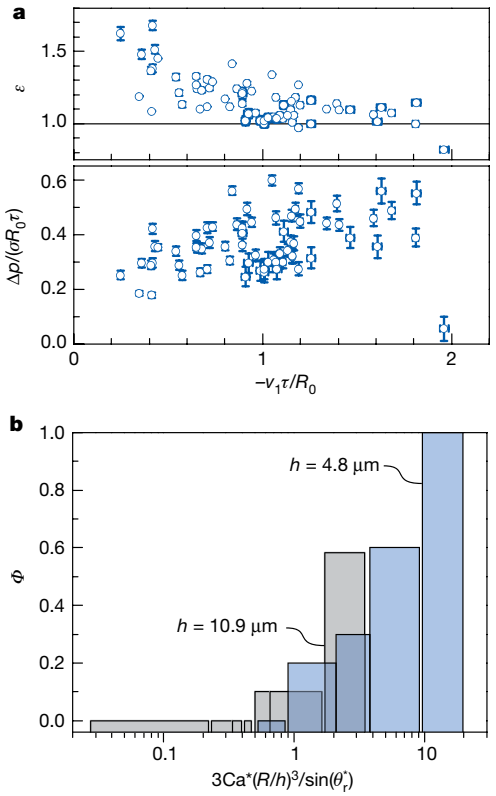


Figure 3 | The effect of microtexture on droplet impact and trampolining dynamics in a low-pressure environment. a, Plots of ε and $\Delta p/(\sigma R_0 \tau)$ versus $-v_1(\tau/R_0)$ for droplets impacting a superhydrophobic surface. Error bars represent uncertainty of the measurement. **b**, Plot of the probability of observing a trampolining event (Φ) versus the force ratio $3\mathrm{Ca}^*(R/h)^3/\sin(\theta_r^*)$ for small droplets ($2R_0 < 0.27$ cm) on superhydrophobic surfaces with similar pillar pitches and diameters, but substantially different heights (blue bar, $h = 4.8$ μm; grey bar, $h = 10.9$ μm; each bar represents an average of at least ten trials). The uncertainty of the force ratio $3\mathrm{Ca}^*(R/h)^3/\sin(\theta_r^*)$ ranges from 17% to 22% (blue bar) and 23% to 31% (grey bar) owing to uncertainties in R, h, P, J, θ_r^* and temperature. Surface properties: **a**, same as Fig. 1; **b**, $[d, l, h] = [1.6, 6.5, 10.9]$ μm and $[1.4, 6.5, 4.8]$ μm.

droplet mass and σ is surface tension. The data indicate that the impact phase, where inertia is important ($-v_1\tau/R_0 \approx 1$), is largely unaffected by pressure; the opposite is true for the recoil dynamics. To better illustrate this, Fig. 2c, d presents plots of droplet–substrate contact radius R as a function of t for low- and standard-pressure cases, and the associated image sequences (see Supplementary Video 4). These plots show that the spreading of the contact line is largely unaffected by pressure, whereas the recession of the contact line changes profoundly (see Fig. 2d). Under standard pressure, during the time $\Delta t_r/\tau$ (shaded yellow region in Fig. 2d), where Δt_r is the duration of the contact-line recession interval, R decreases linearly (slope of -0.6) as a result of inertial and capillary forces balancing each other, and the rate of decrease is consistent with that estimated from a previous retraction model[24]. In the low-pressure case, for the same time region, R decreases parabolically, that is, in a continuously accelerating fashion. The acceleration of the contact-line recession estimated from Fig. 2d is $-51R_0/\tau^2$, which is 22 times greater (in magnitude) than the acceleration experienced by the droplet overall during the corresponding receding/recoil phase (estimated from Fig. 2b to be $2.3R_0/\tau^2$). Thus, this fast acceleration of the contact line acts on the droplet only near the substrate.

The plot of the restitution coefficient $\varepsilon = -v_2/v_1$ (the ratio of outgoing and incoming droplet velocities) as a function of v_1 in Fig. 3a makes it clear that low-pressure conditions increase ε beyond unity for a wide range of impact velocities studied here. The data reveal an inverse dependence of ε on $-v_1$, and that a balance between added momentum Δp and dissipation (and therefore $\varepsilon \approx 1$) is achieved at $-v_1 \approx R_0/\tau$. In Fig. 3a, experimentally determined values of Δp are plotted against v_1. With this data, and by modelling the droplet impact process as a partially inelastic collision that gives the net change in momentum as a result of droplet recoiling (see Methods section 'Inelastic collision model' and Extended Data Figs 3 and 4), we can quantify the overpressure under the droplet and the resultant net force f with the approximation, $\Delta p = \int_0^{\Delta t_r} f\,\mathrm{d}t \approx \overline{f}\Delta t_r$, where Δt_r is the time during which a net force is acting on the droplet (see Fig. 2d) and $\overline{f}$ is the average force over Δt_r. Substituting appropriate values for $-v_1 = 0.6R_0/\tau$ (the impact velocity used in Fig. 2c, d, $\Delta t_r/\tau = 0.16$) yields $\overline{f} \approx 2.2\sigma R_0$. Therefore, for a typical impact velocity occurring during a trampolining event, the average force acting on the droplet during the receding phase, which results in momentum transfer in the y direction, is estimated to be

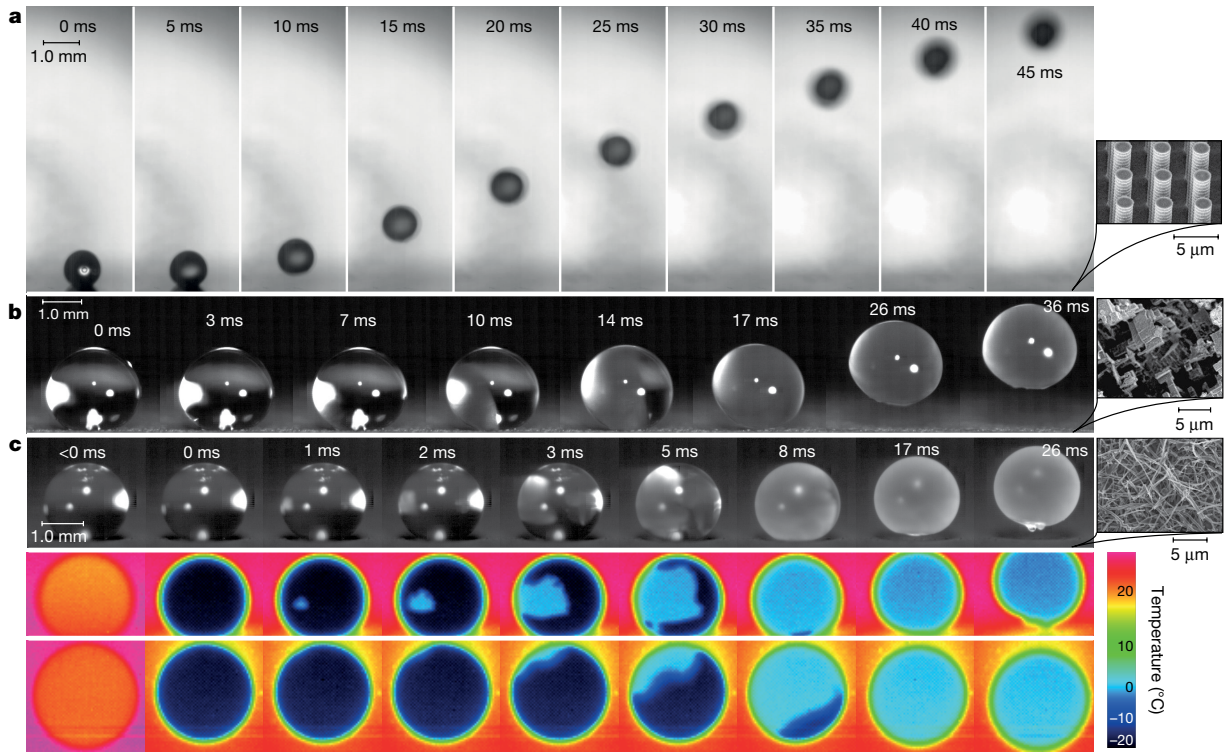


Figure 4 | Freezing can trigger spontaneous droplet launching from a wide range of materials and microtextures. a–**c**, Image sequences showing water droplets solidifying on, and launching from, superhydrophobic surfaces in an environment at standard temperature with low-pressure and low-humidity conditions. The surfaces used were: **a**, silicon micropillar ($[d, l, h] = [2.0, 4.6, 13.5]$ μm; see Supplementary Video 7 for more details); **b**, etched aluminium (Supplementary Video 8); **c**, fluoropolymer–carbon-nanofibre composite (Supplementary Video 9). Micrographs of the three surfaces are given on the right. In **c**, thermographic image sequences are also shown, which are synchronized with the above optical image sequence from side-view (middle row of images) and top-view (bottom row of images) perspectives.

greater than that due to surface tension. The corresponding average overpressure, with respect to the maximum droplet–substrate contact area $R_{max} = 0.63R_0$ (Extended Data Fig. 5), is $\Delta\overline{P} \approx 0.9(2\sigma / R_0)$—a fraction of the Laplace pressure within the droplet.

To determine the physical origin of the force f, it is necessary to consider the interplay between the dynamics of vapour flow and the superhydrophobic surface texture (see Methods section 'Droplet vaporization' and Extended Data Fig. 6). At low pressure and humidity, the liquid droplet has a relatively high vaporization flux J, and the mean free path length λ is relatively high compared to the characteristic length scales of the micropillar array, so slip effects become important (Knudsen number, $Kn = \lambda/h \approx 1$)[25]. If the vapour flux emitted from the droplet into the micropillar array is high enough and the micropillar height small enough, then the vapour will not be able to drain easily through the surface texture (despite its open structure) and an overpressure will result. At vaporization equilibrium, which occurs very rapidly with respect to impact dynamics (see Methods section 'Droplet vaporization'), the quasi-steady-state condition before recession takes place is represented by a balance between the internal pressure of the droplet ($2\sigma/R_0$) and a resisting pressure due to viscosity. We estimate the magnitude of the overpressure by balancing the pressure gradient driving vapour drainage with the viscous stress resisting it[25,26], which, with $R \approx R_0$, is $\Delta P \approx 6\mathrm{Ca}^*(R/h)^3(2\sigma/R)$; here the modified capillary number is defined as $\mathrm{Ca}^* = J\mu/(4F\rho_v\sigma)$ and depends on the vapour viscosity μ, a slip factor F that depends on Kn and wall geometry, and the vapour density ρ_v (ref. 27). Substituting appropriate values yields $\Delta P \approx 4.7(2\sigma/R)$ (ref. 27). The overpressure underneath the droplet that drives the trampolining behaviour varies between this value of ΔP and approximately zero (at the droplet periphery); we estimate it to be the average of those two values: $\Delta\overline{P} \approx 2.3(2\sigma / R)$. This estimate of overpressure is consistent with the result obtained from the inelastic collision model, $\Delta\overline{P} \approx 0.9(2\sigma / R_0)$.

For the trampolining process to self-initiate, the overpressure due to vaporization must overcome substrate adhesion, which is estimated by the force ratio $3\mathrm{Ca}^*(R/h)^3 / \sin(\theta_r^*)$, where θ_r^* is the apparent receding contact angle of the droplet on the superhydrophobic surface. So, for a fixed set of conditions, one can readily satisfy the above criterion by changing R/h. Such an analysis is a conservative estimate and is more accurate for a highly porous surface texture (as most superhydrophobic textures are[28]), which satisfies the condition $dh / l^2 \ll 1$. From a scaling perspective, this criterion suggests that h should be much less than R but still greater than λ: $\lambda < h \ll R$ (see Methods section 'Droplet vaporization'). Whether or not sufficient overpressure was achieved to induce spontaneous levitation in the low-pressure case is shown in Fig. 3b, which plots the probability of observing a trampoline event Φ versus $3\mathrm{Ca}^*(R/h)^3 / \sin(\theta_r^*)$. We used two superhydrophobic micropillar surfaces with relatively constant pillar diameters and pitches, but with substantially different heights, and kept the droplet sizes $2R_0 < 0.27$ cm to minimize gravitational effects. If the droplets are too large, then they are likely to oscillate with a frequency that does not correspond to their travel time in the air, which results in the droplets impacting onto the substrate in an oblate condition. This type of droplet impact can result in $\varepsilon < 1$ in spite of the fact that the droplet is vaporizing (see Supplementary Video 1 and Methods section 'Droplet oscillations'). Although there are inherent experimental deviations, $1 \ll 3\mathrm{Ca}^*(R/h)^3 / \sin(\theta_r^*)$ is a good predictor of whether or not spontaneous droplet trampolining dynamics will occur. The data also show that with shorter pillar heights, one can access a regime where trampolining dynamics should occur practically every time ($\Phi = 1$) for a given droplet size. This observation makes it possible to design a device that generates continuous mechanical motion by coupling the droplet to a cantilever (see Methods section 'Cantilever', Supplementary Video 5 and Extended Data Fig. 7).

Because droplet trampolining dynamics rely on the combined effect of droplet vaporization and substrate–liquid repellence, it would be interesting to explore whether liquids with higher vapour pressure and typically low surface tension can trampoline as well. Although achieving

repellence for that class of liquids is outside the scope of this study because it requires re-entrant superhygrophobic surface features[29,30]. Supplementary Video 6 documents relevant behaviour: a water–acetone droplet initially residing on a standard silicon micropillar superhydrophobic surface spontaneously levitates at a pressure similar to that when water droplets levitate, but exhibits only a few bounces and no sustained trampolining. This lack of trampolining is because the droplet required a relatively large vaporization rate to overcome adhesion with the present substrate, so it transitioned promptly to a Leidenfrost state.

A very noticeable effect of the strong vaporization is the high degree of cooling experienced by the droplets in contact with the surface, which can even cause them to solidify in a recalescent manner (from a supercooled state) in their room-temperature (20–25 °C) environment. This effect is made possible by the relatively high vaporization flux and poor thermal-transport properties of liquid water (see Methods section 'Droplet freezing'), and is consistent with previous work[21] that has shown that evaporative cooling alone can induce a spontaneous recalescent freezing at the free surface of a sessile droplet at the same supercooled temperature as the flow. Furthermore, it is has been shown[20] that the rapid recalescent partial solidification of supercooled sessile droplets is accompanied by a sudden increase of the droplet temperature to its equilibrium value for freezing (0 °C); because the environment is severely undersaturated with respect to water vapour at this temperature, a sudden (explosive) increase in vaporization from the droplet surface occurs. These effects manifest themselves in the present work by rapidly increasing the overpressure under the droplet, which has substantial implications for the ensuing droplet dynamics with respect to ice levitation.

As shown in Fig. 4a and Supplementary Video 7, a sudden overpressure increase between the droplet and substrate, owing to increased evaporation as a result of droplet freezing, is capable of launching it away from the surface at the formative state of icing (see Extended Data Table 2). The phenomena reported in this study are inherent to droplet interactions with textured superhydrophobic surfaces. To underpin this statement, we also demonstrated spontaneous ice levitation on metallic (aluminium) surfaces (see Supplementary Video 8). Figure 4b shows an image sequence of a water droplet freezing and self-levitating from a superhydrophobic, etched aluminium substrate (see inset for a micrograph of the surface), demonstrating a behaviour identical to that of the silicon pillar surface of Fig. 4a. Finally, Fig. 4c shows an image sequence of freezing-driven levitation for a droplet initially in contact with a superhydrophobic, polymer nanocomposite surface (see inset for a micrograph of the fluoropolymer–carbon-nanofibre surface; see Supplementary Video 9 for a video of the image sequence). To quantify the temperature field during the recalescence phenomenon as it triggers ice levitation, we recorded a synchronized thermographic image sequence from side-view and top-view perspectives showing how the temperature of the droplet evolves throughout the freezing process. Initially the droplet is at room temperature; it becomes supercooled owing to vaporization; it then undergoes recalescent freezing (lasting approximately 10 ms), which results in a temperature increase as freezing rapidly spreads along the droplet surface, and triggers its almost immediate levitation. This unexpected ice-levitation process stems from the physics of recalescent freezing, in which the ice nucleation from a supercooled state, the freezing front propagation and the associated vaporization increase are all directly related to the removal of the as-formed ice from a superhydrophobic surface.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 29 May; accepted 10 September 2015.

1. Wisdom, K. M. *et al.* Self-cleaning of superhydrophobic surfaces by self-propelled jumping condensate. *Proc. Natl Acad. Sci. USA* **110,** 7992–7997 (2013).
2. Deng, X., Mammen, L., Butt, H.-J. & Vollmer, D. Candle soot as a template for a transparent robust superamphiphobic coating. *Science* **335,** 67–70 (2012).
3. Schutzius, T. M. *et al.* On the physics of icing and the rational design of surfaces with extraordinary icephobicity. *Langmuir* **31,** 4807–4821 (2015).
4. Boreyko, J. B. & Collier, C. P. Delayed frost growth on jumping-drop superhydrophobic surfaces. *ACS Nano* **7,** 1618–1627 (2013).
5. Bird, J. C., Dhiman, R., Kwon, H.-M. & Varanasi, K. K. Reducing the contact time of a bouncing drop. *Nature* **503,** 385–388 (2013).
6. Liu, Y. *et al.* Pancake bouncing on superhydrophobic surfaces. *Nature Phys.* **10,** 515–519 (2014).
7. Boreyko, J. & Chen, C.-H. Self-propelled dropwise condensate on superhydrophobic surfaces. *Phys. Rev. Lett.* **103,** 184501 (2009).
8. Hou, Y., Yu, M., Chen, X., Wang, Z. & Yao, S. Recurrent filmwise and dropwise condensation on a beetle mimetic surface. *ACS Nano* **9,** 71–81 (2015).
9. Maitra, T. *et al.* On the nanoengineering of superhydrophobic and impalement resistant surface textures below the freezing temperature. *Nano Lett.* **14,** 172–182 (2014).
10. Jung, Y. C. & Bhushan, B. Wetting behaviour during evaporation and condensation of water microdroplets on superhydrophobic patterned surfaces. *J. Microsc.* **229,** 127–140 (2008).
11. Na, B. & Webb, R. L. A fundamental understanding of factors affecting frost nucleation. *Int. J. Heat Mass Transfer* **46,** 3797–3808 (2003).
12. de Ruiter, J., Lagraauw, R., van den Ende, D. & Mugele, F. Wettability-independent bouncing on flat surfaces mediated by thin air films. *Nature Phys.* **11,** 48–53 (2015).
13. Richard, D., Clanet, C. & Quéré, D. Surface phenomena: contact time of a bouncing drop. *Nature* **417,** 811 (2002).
14. Bartolo, D. *et al.* Bouncing or sticky droplets: impalement transitions on superhydrophobic micropatterned surfaces. *Europhys. Lett.* **74,** 299–305 (2006).
15. Eberle, P., Tiwari, M. K., Maitra, T. & Poulikakos, D. Rational nanostructuring of surfaces for extraordinary icephobicity. *Nanoscale* **6,** 4874–4881 (2014).
16. Enright, R., Miljkovic, N., Al-Obeidi, A., Thompson, C. V. & Wang, E. N. Condensation on superhydrophobic surfaces: the role of local energy barriers and structure length scale. *Langmuir* **28,** 14424–14432 (2012).
17. Rykaczewski, K. *et al.* How nanorough is rough enough to make a surface superhydrophobic during water condensation? *Soft Matter* **8,** 8786–8794 (2012).
18. Jung, S. *et al.* Are superhydrophobic surfaces best for icephobicity? *Langmuir* **27,** 3059–3066 (2011).
19. Richard, D. & Quéré, D. Bouncing water drops. *Europhys. Lett.* **50,** 769–775 (2000).
20. Jung, S., Tiwari, M. K. & Poulikakos, D. Frost halos from supercooled water droplets. *Proc. Natl Acad. Sci. USA* **109,** 16073–16078 (2012).
21. Jung, S., Tiwari, M. K., Doan, N. V. & Poulikakos, D. Mechanism of supercooled droplet freezing on surfaces. *Nature Commun.* **3,** 615 (2012).
22. Leidenfrost, J. G. *De Aquae Communis Nonnullis Qualitatibus Tractatus* (Duisburg, 1756); Wares, C. On the fixation of water in diverse fire. *Int. J. Heat Mass Transfer* **9,** 1153–1166 (1966) [transl.].
23. Arpaci, V. S. & Larsen, P. S. *Convection Heat Transfer* 187–190 (Prentice Hall, 1984).
24. Bartolo, D., Josserand, C. & Bonn, D. Retraction dynamics of aqueous drops upon impact on non-wetting surfaces. *J. Fluid Mech.* **545,** 329–338 (2005).
25. Squires, T. M. & Quake, S. R. Microfluidics: fluid physics at the nanoliter scale. *Rev. Mod. Phys.* **77,** 977–1026 (2005).
26. de Gennes, P.-G., Brochard-Wyart, F. & Quéré, D. *Capillarity and Wetting Phenomena: Drops, Bubbles, Pearls, Waves* Ch. 5 (Springer, 2004).
27. Brown, G. P., DiNardo, A., Cheng, G. K. & Sherwood, T. K. The flow of gases in pipes at low pressures. *J. Appl. Phys.* **17,** 802–813 (1946).
28. Ma, M. & Hill, R. M. Superhydrophobic surfaces. *Curr. Opin. Colloid Interf. Sci.* **11,** 193–202 (2006).
29. Tuteja, A. *et al.* Designing superoleophobic surfaces. *Science* **318,** 1618–1622 (2007).
30. Liu, T. "Leo" & Kim, C.-J. "CJ". Turning a surface superrepellent even to completely wetting liquids. *Science* **346,** 1096–1100 (2014).

Supplementary Information is available in the online version of the paper.

Acknowledgements T.M.S. acknowledges the ETH Zurich Postdoctoral Fellowship Program and the Marie Curie Actions for People COFUND programme (FEL-14 13-1). Partial support of the Swiss National Science Foundation under grant number 200021_135479 is also acknowledged. We thank L. J. Yi for his participation in the trampoline experiment, U. Drechsler for advice on surface fabrication and J. Vidic and B. Kramer for assistance in chamber construction.

Author Contributions T.M.S., S.J. and D.P. conceived the project and planned the experiments. T.M., G.G. and T.M.S. fabricated the samples. T.M.S., S.J., G.G. and M.K. carried out the experiments. T.M.S., S.J. and D.P. analysed the data and developed the theoretical analysis. T.M.S., S.J. and D.P. wrote the paper. All authors proofread the paper, made comments and approved the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to D.P. (dpoulikakos@ethz.ch).

METHODS

Droplet trampolining and imaging. The environmental chamber consisted of an aluminium-based pressure vessel connected to a vacuum pump, a pressure sensor and a pressurized nitrogen reservoir (see Extended Data Fig. 8). Temperature and humidity sensors were located at the centre of the vessel. Two transparent PMMA (poly(methyl methacrylate)) windows, one fitted on the front and one on the rear part of the vessel, facilitated the use of a high-speed optical visualization in combination with backlighting, which is arranged in line with the lens of the camera, to visualize droplet trampolining. For each experiment, a 1−10-μl deionized water droplet was placed on the sample surface, and then the environmental pressure at the inner volume of the vessel was reduced using a valve connected to the vacuum pump at a rate of -0.1 bar s^{-1}. Throughout the experiments, the ambient pressure and temperature inside the vessel were kept constant at chosen values (approximately 0.01 bar and 20–30 °C, respectively). Prior to initiating an experiment (that is, reducing environmental pressure), the entire pressure chamber was purged with nitrogen to ensure a dry environment. The droplet dynamics were captured by high-speed video recording at 500–50,000 frames per second. The same procedure was followed for the droplet–cantilever experiments. The steel cantilever had a width, length and thickness of 1.2 cm, 5.1 cm and 80 μm, respectively.

To understand whether or not a droplet will trampoline on a given surface, we provide the following example. For a surface with $h \approx 10$ μm, one can obtain information on the droplet sizes for which trampolining will occur from Fig. 3b. We estimate the size of the droplet–substrate contact radius R at which trampolining is expected to occur. For a surface with $h = 10.9$ μm, $\theta_r^* = 154°$ (Extended Data Table 1) and a vaporization flux of $J = 6.9 \times 10^{-4}$ g cm^{-2} s^{-1} (Extended Data Fig. 6), to achieve $3\mathrm{Ca}^*(R/h)^3/\sin(\theta_r^*) \approx 3$—the condition where trampolining is more likely to occur for $2R_0 < 0.27$ cm—then $\mathrm{Ca}^* = 1.5 \times 10^{-6}$ and $R/h = 67$. The latter condition leads to $R = 0.73$ mm. For such a contact radius, the typical value of the initial droplet radius is then $R_0 = 1.34$ mm.

Vaporization rate. We determined the vaporization flux of a droplet in contact with a superhydrophobic silicon micropillar surface ($[d, l, h] = [1.4, 6.5, 18.2]$ μm) as a function of environmental pressure in a dry environment by determining the cross-sectional area of the droplet in time using ImageJ software (see Methods section 'Droplet vaporization'). For the low-pressure conditions of this study, we determined the vaporization flux for millimetre-scale droplets on a superhydrophobic surface with a solid/air fraction of $\phi = 0.04$ to be 0.69 mg cm^{-2} s^{-1}.

Silicon micropillar surface. A polished boron-doped p-type (100) silicon wafer with areal dimensions of 1.5×1.5 cm^2 and a thickness of 500 ± 25 μm was used as the substrate. Photolithography was performed using an AZ 1505 and an AZ 6612 positive photoresist with a Karl Suss MA6 mask aligner; material removal was performed using inductive-coupling plasma etching with SF_6/C_4F_4 (Bosch Process in Alcatel AMS 200 machine) to fabricate a regular and well defined micropillar surface structure. To lower the surface energy of the textured surface, rendering it hydrophobic, a layer of 1H,1H,2H,2H-perfluorodecyltrichlorosilane was applied by liquid-phase self-assembly. For full details on the characteristics of these surfaces (for example, geometry, wettability and so on), see Extended Data Table 1.

Etched aluminium surface. To generate the superhydrophobic aluminium surface, we used a procedure inspired by ref. 31. An aluminium substrate with areal dimensions of 2×2 cm^2 (weight fractions: Al, 99.58%; Si, 0.1%; Fe, 0.12%; Cu, 0.03%; Mn, 0.02%; Mg, 0.02%; Zn, 0.03%; Ti, 0.02%; Ga, 0.03%; and V, 0.05%. Bronmetal, AW1085 from Bronmetal) was initially cleaned under sonication in acetone, isopropyl alcohol and deionized water for 10 min each. Subsequently, to remove the native oxide layer, the substrate was treated with a 1 wt% sodium hydroxide solution for 10 min. Thereafter, the aluminium substrate was etched with a 1 M ferric chloride solution for 25 min at 50 °C. During the etching step, the aluminium substrate was cleaned with isopropyl alcohol for 2–3 min to avoid precipitation of ferric hydroxide on the surface. Finally, to impart hydrophobicity, the etched aluminium surface was treated with a 1.43 mM solution of trichloro-1H,1H,2H,2H-perfluorodecylsilane in *n*-hexane for 2 h followed by heating for 45 min at 120 °C. For details on the wettability and morphological characteristics of these surfaces, see Extended Data Table 1.

Polymer nanocomposite. To generate the superhydrophobic polymer nanocomposite coating, we used a procedure inspired by ref. 32. To begin, an aqueous fluoroacrylic copolymer dispersion (PMC, 20 wt% in water; Capstone ST-100, DuPont) was diluted with acetic acid (0.4 wt% PMC in acetic acid/water); separately, a suspension of carbon nanofibre particles (CNF; diameter ≈ 100 nm and length ≈ 20–200 μm, >98% carbon basis; Sigma Aldrich) in acetic acid was generated (CNF 2.0 wt% in acetic acid). Both solutions were separately subjected to 30 min of ultrasonic probe sonication (Vibracell VCX 130, 130 W, 20 kHz). The CNF and PMC dispersions were then combined and mechanically mixed at room temperature to generate the final dispersion. The final solid weight ratio of PMC to CNF was 1:5; the total solid concentration (CNF + PMC) in the dispersion is 1.1 wt%. The PMC–CNF dispersion was then probe-sonicated for 30 min. Finally, the dispersion was spray-deposited with a siphon-feed air brush onto standard glass slides from a distance of approximately 10 cm and the coatings were placed on a hot plate (approximately 100 °C) for several minutes to facilitate the removal of residual solvents. The morphological and wettability characteristics of the surface are shown in Extended Data Table 1.

Surface characterization. For surface morphology characterization, we used a scanning electron microscope (Zeiss ULTRA 55); we applied no conductive coatings to facilitate imaging. We performed advancing and receding contact angle measurements with a backlit image acquisition setup (goniometer) consisting of a syringe pump (for dispensing and withdrawing volume of the droplet on the substrate) and a detector (Thorlabs, DCC1645C) affixed with a standard zoom lens (Thorlabs, MVL7000) for the purposes of droplet visualization.

Modelling droplet trampolining. When the droplet is in contact with the substrate, the droplet trampolining behaviour can be described as a standard mass–spring–damper (MSD) system with a forcing function as depicted in Extended Data Fig. 1 (for further details, see Supplementary Video 2 and Fig. 1). When the droplet is not in contact with the substrate, it is governed by projectile motion. Therefore, we describe the dynamics of this entire system as

$$\begin{aligned} y &= y_0 + v_0 t - \frac{1}{2} g t^2 \quad \text{if } y \geq 0 \\ m\frac{\mathrm{d}^2 y}{\mathrm{d}t^2} + c\frac{\mathrm{d}y}{\mathrm{d}t} + f_k(y) &= f(t) - mg \quad \text{if } y < 0 \end{aligned} \tag{1}$$

where m is the droplet mass, y is the vertical position of the cross-sectional centroid of the droplet, t is time, c is the damping constant, $f_k(y)$ is the force due to the 'stiffness' of the droplet, $f(t)$ is a forcing function, v is velocity and g is gravitational acceleration (zero subscripts denote initial values). Selecting appropriate scales, we transform equation (1) into non-dimensional form

$$\begin{aligned} y^* &= y_0^* + v_0^* t^* - \frac{1}{2}\mathrm{Bo}(t^*)^2 \quad \text{if } y^* \geq 0 \\ \frac{\mathrm{d}^2 y^*}{\mathrm{d}t^{*2}} + 2\zeta\frac{\mathrm{d}y^*}{\mathrm{d}t^*} + f_k^* &= f^*(t^*) - \mathrm{Bo} \quad \text{if } y^* < 0 \end{aligned} \tag{2}$$

where $\mathrm{Bo} = mg/\sigma R_0$ is the gravitational Bond number, R_0 is the initial droplet radius, σ is surface tension, $2\zeta = c/\sqrt{\sigma m}$ is the damping ratio and $\tau = \sqrt{m/\sigma}$ is the inertial-capillary timescale. We write the dimensionless variables (indicated by asterisks) as

$$\begin{aligned} y_0 &= R_0 y^* \\ t &= \tau t^* \\ v &= v^* R_0/\tau \\ f_k &= \sigma R_0 f_k^* \\ f(t) &= \sigma R_0 f^*(t^*) \end{aligned}$$

We distinguish three cases: $0 < \zeta < 1$ (under-damped), $\zeta = 1$ (critically damped) and $\zeta > 1$ (over-damped). In the traditional MSD system, the critically damped and over-damped cases have no harmonic component in their solutions and, therefore, they are not relevant to the problem here. In our study, we estimate that $\zeta \approx 0.11$ (see Methods section 'Inelastic collision model'), indicating that we have an under-damped condition; this is the value used to illustrate the phenomenon for the duration of this section.

As was mentioned above, each trampolining cycle is governed by both MSD dynamics ($y^* < 0$) and projectile motion ($y^* \geq 0$); however, it is the forcing-function frequency matching the MSD frequency—and the fact that this force only acts when $y^* < 0$—that is responsible for the trampolining behaviour. To fully represent the droplet trampolining as a MSD system, a representative model of droplet stiffness needs to be developed.

We develop a comprehensive model for stiffness that takes into account the change in shape of the droplet as it deforms. Previous approaches have provided simplified and rigorous descriptions of droplet elasticity[19,33–36]. Our aim is to derive a simple model describing the force due to surface tension, as a function of droplet deformation in the direction parallel to droplet transport, where the force causing deformation is due to inertia.

If one assumes that at the moment of maximum deformation the droplet takes the shape of an incompressible flat cylinder of radius W, height H and volume V, then f_k^* is

$$f_k^* = 2\pi\left[\sqrt{\frac{1}{(y^*+1)}} - \frac{1}{(y^*+1)^2}\right]$$

where $y = (V/(2\pi))^{1/3}y^*$, which we refer to as the 'cylinder stiffness model'. For the theoretical cylinder geometry, we use an effective radius R_0 from the volume relation $V = 2\pi R_0^3$, which leads to $y = R_0 y^*$ ($2W = H = 2R_0$ in the undeformed case). The above equation implies that f_k^* depends nonlinearly on y^*, indicating that the droplet behaves like a spring with variable stiffness.

The final variables required to solve equation (2) are the initial velocity v_0, the initial droplet position y_0 (typically zero) and the expression for $f(t)$. For the case where no force is applied, that is, a droplet is impacting a non-wetting surface with negligible force generation from evaporation effects, $f(t) = 0$. If $f(t) \neq 0$, that is, a droplet is impacting a non-wetting surface under low-pressure and humidity conditions that drive force-generating evaporation, then we define the force as the piecewise-continuous function

$$f(t) = \begin{cases} 0 & \text{if } \frac{dy}{dt} \leq 0\,, \quad y(t) < 0 \\ \overline{f} & \text{if } \frac{dy}{dt} > 0\,, \quad y(t) < 0 \end{cases}$$

where $\overline{f}$ is the average value of the force applied to the droplet during the recoiling stage of the droplet impact process. The above equation states that a force is only being applied to the droplet, in a positive direction, once the droplet is in the recoiling stage of the impact process.

Extended Data Figure 2a presents a plot of a full solution for equation (2) compared against experimental data. It is clear that there is a marked deviation between the experimental and theoretical results in terms of projectile motion for the first levitation sequence, which is due to the choice of $v_0^*(t^* = 0)$: in the experimental case, in the first few bounces, although the droplet does not reach the theoretical maximum heights, it stores a substantial amount of energy in the form of inherent capillary waves from its motion, which ultimately contribute to the lower values of $y_{\min}^*$. This is demonstrated by the inset of Extended Data Fig. 2a, which shows an image of an elongated water droplet at the moment of impact at $t/\tau = 2.2$. This same behaviour is true for successive impact sequences where $t/\tau < 20$. (Fig. 1 shows images of elongated water droplets at the moment of impact that correspond to the sequence shown in Extended Data Fig. 2a). For higher droplet impact velocities, the role of capillary waves in storing kinetic energy is reduced, as shown at later bounces in Extended Data Fig. 2a, where the droplet bouncing heights are noticeably higher and the theoretical and experimental cases match very closely. Extended Data Fig. 2b shows the force profiles required to induce trampolining behaviour for the theoretical case in Extended Data Fig. 2a. The magnitude of the average force acting on the droplet during the recoiling stage of droplet impact is of the same scale as the force due to surface tension; this is result is further validated in Methods section 'Droplet vaporization'.

By capturing the magnitude, direction and timing of force application on a simple, under-damped MSD system with a spring that has variable stiffness, we have reproduced the experimentally observed behaviour relatively well considering the model simplicity, particularly the periodic, quasi-steady-state condition. This outcome underpins the claim that the trampolining behaviour of the droplet can be described by an under-damped, forced, MSD–projectile system operating at resonance.

Droplet mass loss. One hypothesis for a droplet impact event resulting in $\varepsilon = -v_1/v_2 > 1$ (ratio of recoiling and impacting velocities) that must be considered is that the droplet loses an appreciable amount of mass during the time of the impact event—that is, the impacting masses of the droplet in two successive periods are not equal ($m_2 \neq m_1$). Assuming negligible momentum loss due to viscous dissipation in the air, we estimate the expected value of ε due to mass loss to be $\varepsilon \approx m_1/m_2$; therefore, to achieve a typical restitution coefficient of $\varepsilon = 1.24$ ($(m_1 - m_2)/m_1 = 0.2$—that is, the droplet should shed 20% of its mass during one trampolining period. For the experiments performed in this study, we estimate the mass loss of a droplet during a single trampolining cycle as

$$\frac{m_1 - m_2}{m_1} \approx \frac{3J\Delta t}{\rho R_1}$$

where R_1 is the initial (at the start of the trampolining cycle) radius of the droplet, J is the vaporization flux of the droplet, Δt is the time it takes to complete one trampolining cycle and ρ is the density of the droplet. In Methods section 'Droplet vaporization', the vaporization flux of the droplet on a superhydrophobic surface is measured to be $J \approx 6 \times 10^{-4}\,\mathrm{g\,cm^{-2}\,s^{-1}}$; Extended Data Fig. 2 shows that $\Delta t \approx 5\tau$. So, for a water droplet with $R_1 \approx 0.1$ cm ($m_1 = \rho(4/3)\pi R_1^3$), we calculate the inertial-capillary timescale to be $\tau = \sqrt{m/\sigma} = 7.6$ ms, where σ is surface tension ($\rho = 1.0\ \mathrm{g\,cm^{-3}}$, $\sigma = 72\ \mathrm{Dyn\,cm^{-1}}$). Substituting the above values yields $(m_1 - m_2)/m_1 \approx 7 \times 10^{-4}$, that is, the droplet sheds only 0.07% of its mass during a single trampolining cycle, which indicates that the experimentally observed vaporization flux is about 290 times less than that required to induce trampolining by the mass-loss mechanism.

Inelastic collision model. For the inelastic collision model presented within the main text (Fig. 3), one of the most important unknown parameters is the damping ratio ζ, which is a measure of energy dissipation as a function of impact velocity. We estimate an average value of ζ for the superhydrophobic surface and impact velocities of interest to this study. To place this value of ζ into context, we compare it with values obtained in a previous study[19] for an ultralow-hysteresis superhydrophobic surface, which is a good measure of the absolute reachable lower limit for ζ. We then determine a reasonable estimate of the additional momentum (per cycle) required to sustain droplet trampolining.

The damping ratio can be estimated experimentally by knowing the coefficient of restitution $\varepsilon = -v_2/v_1$, the impact velocity v_1 and the droplet–substrate contact time t_c, and is defined as $\zeta = 0.5(1-\varepsilon)\tau/t_c$, where $\tau \equiv \sqrt{m/\sigma}$ is the inertial-capillary timescale, m is the droplet mass and σ is the coefficient of surface tension. For the case of a droplet impacting a surface with negligible contact angle hysteresis at relatively low impact speeds, Richard and Quéré[19] suggest that the maximum value of restitution should be $\varepsilon = \sqrt{5/6} \approx 0.91$, owing to kinetic energy being converted to vibrations. Because we are tracking the translational motion of the droplet, if kinetic energy is converted into droplet oscillations, then effectively that will appear as a loss energy in terms of lower droplet speed, particularly for the cases where the droplet spends a large amount of time in the air and the oscillations have sufficient time to dissipate.

Extended Data Figure 3a is a plot of ε versus v_1 for two different surfaces: (1) the surface used in this study ($\cos(\theta_r^*) - \cos(\theta_a^*) = 0.14$); and (2) and an ultra-low-hysteresis superhydrophobic surface ($\cos(\theta_r^*) - \cos(\theta_a^*) = 0.02$; ref. 19). (Here θ_r^* and θ_a^* are the receding and advancing contact angles, respectively.) For the surface used in this study, the average value of the restitution coefficient for the range of impact velocities of interest is $\varepsilon = 0.71$—much smaller than the theoretical limit (owing to much larger contact-angle hysteresis). For droplet impact experiments in a low-pressure environment, contact-line pinning is not observed because the dewetting process occurs almost instantaneously with respect to the recoiling of the droplet (see Supplementary Videos 3 and 4); therefore, one should expect that for such a superhydrophobic surface with relatively higher contact-angle hysteresis, the value of ζ that is measured in an environment at standard pressure would be an overestimation for the same experiment in a low-pressure environment (see Extended Data Fig. 3b). Furthermore, because contact-line pinning does not occur in a low-pressure environment, the impact behaviour of a droplet on a superhydrophobic surface should tend towards the ideal case described in ref. 19 ($\varepsilon \approx 0.91$). In this case, if the droplet contact time is the so-called minimum contact time[5,6] ($t_c/\tau = \sqrt{6\pi}/4 \approx 1.09$; Extended Data Fig. 4), which is defined as the lowest-order oscillation period for a spherical droplet, then one should expect $\zeta \approx 0.04$, 40% of the average value determined experimentally for the ambient case ($\zeta \approx 0.11$; see Extended Data Fig. 3b). Using the magnitude/range of ζ estimated from the ambient case and the theoretical considerations, the positive change in momentum as a result of the droplet impact event, Δp, is determined as a function of v_1 (Fig. 3a). For the purposes of conservatively estimating Δp and modelling the process as a MSD system, we chose to use the upper value of $\zeta = 0.11$ throughout the manuscript. Choosing the lower value would result in a pre-factor adjustment to the estimation of Δp; however, it would not change the order-of-magnitude estimate.

We write an approximate definition of Δp as

$$\frac{\Delta p}{\sigma R_0 \tau} = v_1 \frac{\tau}{R_0}\left[(1-\varepsilon) - 2\zeta\left(\frac{t_c}{\tau}\right)\right]$$

where $\Delta p = \int_0^{\Delta t_r} f\,dt$ is the positive change in momentum (from a force f being applied while the contact line recedes), t_c is the droplet–substrate contact time and Δt_r is the time where a net force is acting on the droplet (see Fig. 2d). By knowing ε, ζ and t_c as functions of v_1 for droplet impact experiments performed in a dry low-pressure environment, we determine Δp—as is done in Fig. 3a—and therefore estimate f.

We determined Δp as a function of v_1, and then an average force, by re-writing the definition of Δp as $\Delta p = \int_0^{\Delta t_r} f\,dt \approx \overline{f}\,\Delta t_r$. Substituting appropriate values ($-v_1 = 0.6R_0/\tau$, $\Delta p = 0.35\sigma R_0 \tau$, $\Delta t_r = 0.16\tau$) yields $\overline{f} \approx 2.2\sigma R_0$. (It is instructive to compare the magnitude of this force with the MSD model in Extended Data Fig. 2, which estimated that the average force acting on the droplet during a similar recoil phase of the droplet impact process ($-v_1 = 0.5R_0/\tau$) was $\overline{f} \approx 0.8\sigma R_0$ for $\Delta t_r = 0.41\tau$;

therefore, $\Delta p = 0.33\sigma R_0\tau$, which compares well with the above value.) This force can be transformed into pressure by using the maximum droplet–substrate contact area ($\Delta\overline{P} = \overline{f}\ /\ (\pi R_{max}^2)$); the desired parameter, R_{max}, is shown in Extended Data Fig. 5 as a function of v_1, with its behaviour being consistent with that reported previously[37]. Substituting appropriate values yields $\Delta\overline{P} = (2.2\sigma R_0)\ /\ (\pi R_{max}^2)$ or $\Delta\overline{P} \approx 0.9(2\sigma\ /\ R_0)(\overline{f}\ \approx 2.2\sigma R_0$ and $R_{max} = 0.63R_0$), which represents a fraction of the Laplace pressure in the droplet.

Droplet vaporization. For droplets evaporating in a low-pressure, low-humidity environment when in contact with a superhydrophobic surface, it is possible to estimate the rate of evaporation by measuring how the cross-sectional area changes in time. If we assume that the cross-sectional area measured is equivalent to that of a sphere cross-section (negligible gravitational effects, $\mathrm{Bo} = mg\ /\ (R_0\sigma) \ll 1$; ideally non-wetting, $\theta^* \approx 180°$), we determine an effective radius as a function of time, $R_e(t)$. (Here θ^* is the apparent contact angle that the droplet forms with the substrate.) The rate of change of radius is then used to determine the vaporization flux. Using the liquid density ρ, we obtain the vaporization flux

$$-J \approx \rho\frac{\mathrm{d}R_e}{\mathrm{d}t}$$

Extended Data Figure 6 shows the values of J for millimetre-scale water droplets on a superhydrophobic surface with a similar wetting fraction to the surfaces used throughout this study ($\phi = 0.04$), as a function of environmental pressure. It is clear that for environmental pressures below 0.1 bar, the vaporization flux increases markedly. We also found that for the early stage of droplet evaporation (up to about 30 s), the vaporization flux of the droplets does not change appreciably, which justifies its treatment as a constant value for the subsequent analysis.

Now that the vaporization flux has been estimated, we consider the behaviour of vapour flowing into the micropillar array. The Knudsen number determines whether or not a gas can be treated as a continuous medium, which is important for this problem owing to the low-pressure environment and the small feature sizes of the micropillar array. It is defined as the ratio of the mean free path λ to a characteristic length scale of the system. Here, because the region of interest is within the micropillar array, the characteristic length scale should relate to the micropillar length scales. For the purposes of understanding the main mechanism of how pressure is distributed beneath the droplet, we choose to treat the flow there as a one-dimensional channel, which is possible if the surface satisfies the criterion $dh\ /\ l^2 \ll 1$, that is, it contains sparsely spaced pillars with small diameters. The most important length scale is the height of the pillars h, which defines the scale of the gap in which the vapour flows; therefore, $\mathrm{Kn} = \lambda/h$. The mean free path is a function of pressure, and is determined by

$$\lambda = \frac{k_B T}{\sqrt{2}\,\pi d_v^2 P}$$

where k_B is the Boltzmann constant, T is the temperature, d_v is the effective diameter of the vapour molecule (Lennard–Jones parameter) and P is the absolute pressure. We estimate the mean free path as $\lambda = 12\,\mu\mathrm{m}$ ($k_B = 1.38 \times 10^{-16}\,\mathrm{erg\,K^{-1}}$; $T = 273\,\mathrm{K}$; $d_v = 2.64 \times 10^{-8}\,\mathrm{cm}$ (ref. 38); $P = 0.01$ bar) and, with a pillar height of $h = 4.8\,\mu\mathrm{m}$, $\mathrm{Kn} = 2.5$. Because the vapour pressure of water under standard conditions is 0.032 bar, even if a region of the chamber is at saturation conditions, the value of Kn computed above should be similar.

For a steady-state, incompressible flow in a one-dimensional channel (of channel height h) in the x direction with slip-velocity boundary conditions, the volumetric flow rate (in units of $\mathrm{cm^2\,s^{-1}}$) is[27,39]

$$Q = -\frac{h^3}{12\mu}\frac{\partial P}{\partial x}\left[1 + 12\left(\left(\frac{2}{\Psi} - 1\right)\frac{\lambda}{h}\right)\right] \quad (3)$$

where $(1 - \Psi)$ represents the fraction of molecules colliding with a wall that are reflected[39]. The second term in the square brackets represents a modification due to slip effects, and it is customary to define

$$F = 1 + 12\mathrm{Kn}\left(\frac{2}{\Psi} - 1\right) \quad (4)$$

For air flowing through round glass tubes and $\mathrm{Kn} < 1.0$, Brown *et al.*[27] suggests that $\Psi = 0.84$. By taking this value and $\mathrm{Kn} = 2.5$ and substituting into equation (4), we see that $F = 43$. Hence, the volumetric flow rate (equation (3)) is $Q = -\frac{h^3}{12\mu}\frac{\partial P}{\partial x}F$. If the pressure drop is linear from the centre of the droplet to the edge, we estimate $\frac{\partial P}{\partial x} \approx -\Delta P\ /\ R$. Substituting and rearranging yields

$$\Delta P \approx \frac{12Q\mu R}{h^3 F} \quad (5)$$

We define Q in terms of J by accounting for the geometry of the droplet and the underlying substrate and defining an average velocity within the channels $\overline{u}_v$. From conservation of mass, we see that $\overline{u}_v = (J\ /\ \rho_v)(R\ /\ (4h))$, where ρ_v is the vapour density (found from the ideal gas law, $\rho_v = P(M\ /\ \overline{R})(1\ /\ T)$, with M denoting the molar mass and $\overline{R}$ the universal gas constant) and R is the contact radius between the droplet and the substrate. The volumetric flow rate (in units of $\mathrm{cm^2\,s^{-1}}$) is then $Q = \overline{u}_v h$; therefore, equation (5) becomes

$$\Delta P \approx \frac{3\mu R^2 J}{h^3 F\rho_v} \quad (6)$$

By assuming that $R \approx R$ and substituting appropriate values ($\mu = 1.0 \times 10^{-4}$ P; $R = 0.050\,\mathrm{cm}$; $J = 6.87 \times 10^{-4}\,\mathrm{g\,cm^{-2}\,s^{-1}}$; $h = 4.8 \times 10^{-4}\,\mathrm{cm}$; $F = 43$; $\rho_v = 7.94 \times 10^{-6}\,\mathrm{g\,cm^{-3}}$), we see that $\Delta P \approx 4.7(2\sigma/R)$. On the basis of this, the average overpressure under the drop is $\Delta\overline{P} \approx 2.3(2\sigma/R)$.This value is markedly larger than the average pressure rectified into the vertical motion of the droplet (see Methods section 'Inelastic collision model', $\Delta\overline{P} \approx 0.9(2\sigma\ /\ R_0)$), which shows that this overpressure is the origin of the force driving the trampolining behaviour.

One important aspect of the behaviour of the droplet is how strongly the pressure difference depends on pillar height and droplet radius. Taking $R \approx R_0$ (contact radius and droplet radius are similar), and non-dimensionalizing the pressure difference from equation (6) against Laplace pressure, yields

$$\Delta P\left(\frac{R}{2\sigma}\right) \approx \frac{6\mathrm{Ca}}{F}\left(\frac{R}{h}\right)^2 = 6\mathrm{Ca}^*\left(\frac{R}{h}\right)^3 \quad (7)$$

where $\mathrm{Ca} = \frac{\overline{u}_v\mu}{\sigma} = \frac{J\mu}{4\rho_v\sigma}\left(\frac{R}{h}\right)$ and $\mathrm{Ca}^* = \frac{J\mu}{4F\rho_v\sigma}$ represent the capillary and modified capillary numbers for this problem, respectively. This equation shows that for droplet trampolining to occur, for a given droplet size (for example, $R \approx 0.1$ cm), the pillar height should be about two orders of magnitude less than the radius of the droplet to ensure that a sufficient pressure difference can build up. Furthermore, this analysis holds when the vapour flow can be treated as a continuum; therefore, for these specific low-pressure conditions, $h > 12\,\mu\mathrm{m}$ ($\mathrm{Kn} < 1$). So, for a typical millimetre-scale droplet, a pillar height of about 10 μm should be sufficient to promote trampolining behaviour. For this process to be spontaneous, the pressure difference should be sufficient to overcome the force due to adhesion, which for a superhydrophobic surface is $f_c = 2\pi R\sigma\sin(\theta_r^*)$, where θ_r^* is the apparent receding contact angle of a droplet on a surface. By recalling the definition for the pressure difference ($\Delta P \approx 3\mu R^2 J/(h^3 F\rho_v)$), averaging it and projecting it onto the contact area of the surface, we determine the force due to vaporization as

$$f_v = \Delta\overline{P}\pi R^2 \approx \frac{3\pi\mu R^4 J}{2h^3 F\rho_v}$$

By taking the ratio of the two forces, we develop an approximate criterion for the initiation of trampolining behaviour

$$\frac{f_v}{f_c} = \frac{3\mathrm{Ca}^*}{\sin(\theta_r^*)}\left(\frac{R}{h}\right)^3$$

This criterion is similar to equation (7), with the exception of a $1\ /\ (2\sin(\theta_r^*))$ term (for hydrophobic surfaces, $1\ /\ \sin(\theta_r^*) > 1$). So, a simple, order-of-magnitude design rule for trampolining dynamics on superhydrophobic micropillar arrays can be summarized as: for a droplet of a given radius R_0 that forms a contact radius with a substrate of R, $\lambda < h \ll R$.

We use this analysis to determine the so-called drainage timescale τ_d. If this timescale is much smaller than the timescale of droplet impact and recoil τ, then we can treat the drainage process as a steady-state problem and the estimated value of ΔP will be a good approximation. This timescale is estimated as the ratio of the length traversed to the average gas velocity, $\tau_d \approx R\ /\ \overline{u}_v$. Substituting appropriate values yields $\tau_d/\tau = 0.003$; therefore, the drainage process is two orders of magnitude faster than the natural time-scale of the system (impact and recoil). It is also instructive to compare τ_d with the crashing (impacting) time of the droplet, $\tau_v \approx -R_0/v_1$. Substituting appropriate values yields $\tau_d/\tau_v = 0.004$ ($R_0 = 0.1$ cm; $-v_1 = 19.7\,\mathrm{cm\,s^{-1}}$); therefore, the drainage process is over two orders of magnitude faster than the crashing timescale of the system.

Droplet oscillations. If the droplet oscillation and the time that the droplet is in the air are not synchronized, then the droplet can impact the substrate in a (for example) sufficiently oblate condition. This can result in an inefficient transfer of surface energy to kinetic energy, as well as premature dewetting of the substrate, causing the restitution coefficient to be less than unity, as shown in Supplementary Video 1. The premature dewetting results from an enhanced

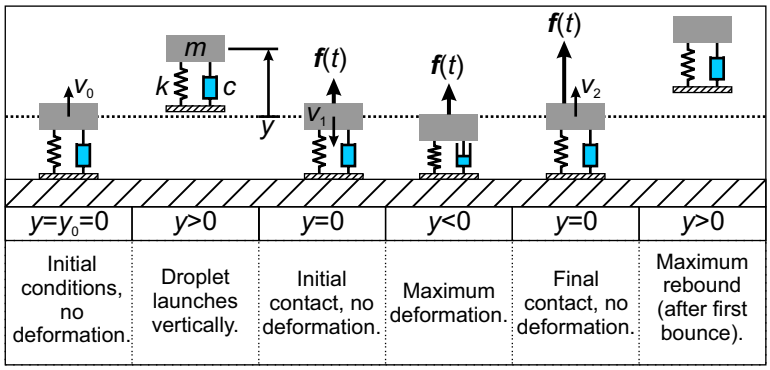
droplet–substrate contact area—causing a relatively high overpressure—which forces the droplet from the surface at the moment of maximum droplet deformation, reminiscent of so-called 'pancake bouncing'[6]. Such behaviour can be achieved by using droplets with diameters that approach the capillary length L_{cap}, because the time that the droplet is in the air is never long enough to fully dissipate its oscillations (Supplementary Video 1). Also, for early trampolining dynamics, adjacent cycles have periods with differing lengths, because the droplet continuously spends more and more time in the air (jumping higher and higher)—as shown by Fig. 1—so a situation where the free oscillation frequency of the droplet, which is a fixed quantity, becomes out of phase with the time the droplet spends in the air may occur.

Large droplets ($h \ll R \approx L_{cap}$) are likely to oscillate with a frequency that does not correspond to the time that the droplet is in the air, resulting in droplets impacting the substrate in an oblate condition. This type of droplet impact can result in $\varepsilon < 1$ in spite of the fact that the droplet is vaporizing; however, the robustness of the trampolining phenomenon is also apparent in Supplementary Video 1: when a droplet impacts with $\varepsilon < 1$ the process is subsequently shown to recover and return to bouncing with $\varepsilon > 1$.

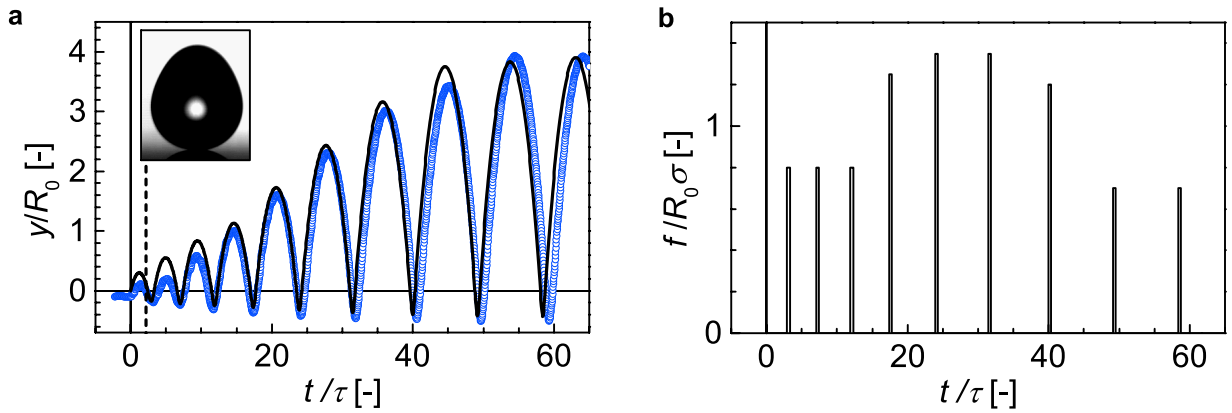
Cantilever. Because the phenomenon of droplet trampolining has a natural frequency (of the order of $1/\tau$) and a predictable force (about $\bar{f}$), we can further quantify and potentially exploit its manifestation for the purposes of inducing continuous motion in a simple device for mechanical power production. To demonstrate this, consider the image sequence in Extended Data Fig. 7a and the associated plot of beam position δ as a function of t in Extended Data Fig. 7b, where a single droplet is attached to a thin, metallic, cantilever beam, and it impacts a superhydrophobic surface in a cyclic manner. This system generates continuous sinusoidal motion for about 400 cycles, after starting spontaneously from rest, compared to about 3 cycles for standard pressure conditions, which required an initiation pulse (see Supplementary Video 5). Localizing and harnessing the trampolining behaviour under the cantilever beam, the power of this phenomenon is visualized in terms of continuous generation of kinetic energy, which drives the cantilever motion well past its natural oscillation, also shown in Supplementary Video 5. Furthermore, Supplementary Video 5 shows a single half-cycle of beam oscillation with high temporal resolution, and captures the marked dewetting behaviour of the droplet, underpinning the above claims.

Droplet freezing. Owing to the large vaporization rates experienced by the droplet, the relatively low thermal conductivity of water and the low liquid–solid wetting fraction ϕ of the surface texture, a large temperature difference can develop across the droplet. We estimate the magnitude of the temperature difference for a water droplet on a silicon micropillar surface using a one-dimensional model, which balances the rate of heat absorbed at the surface of the droplet owing to evaporation, $q \approx 4\pi R_0^2 J \Delta H_{vap}$ (here ΔH_{vap} is the enthalpy of vaporization), with the rates of heat transferred to the interior of the droplet and to the substrate (heat losses to the environment are assumed to be negligible). We estimate the thermal resistance of the droplet, by treating it as a shell with an inner (R_1) and outer (R_2) radius, as $\mathcal{R}_w = (1/R_1 - 1/R_2)/(4\pi k_w)$, and that of the composite air–micropillar region as $\mathcal{R}_i = h/(k_e \pi R^2)$, where k_w is the effective thermal conductivity of the water droplet and the effective thermal conductivity of the air–micropillar region is defined as $k_e = \phi k_{Si} + (1-\phi)k_{air}$, with k_{Si} and k_{air} the thermal conductivities of silicon and air, respectively. Substituting appropriate values yields $k_e = 0.06\,\mathrm{W\,cm^{-1}\,K^{-1}}$ ($\phi = 0.04$, $k_{air} = 2.6 \times 10^{-4}\,\mathrm{W\,cm^{-1}\,K^{-1}}$, $k_{Si} = 1.48\,\mathrm{W\,cm^{-1}\,K^{-1}}$) (ref. 40). We compare the magnitudes of these individual resistances by substituting appropriate values, which yields $\mathcal{R}_w = 133\ \mathrm{K\ W^{-1}}$ and $\mathcal{R}_i = 3\ \mathrm{K\ W^{-1}}$ ($k_w = 6 \times 10^{-3}\,\mathrm{W\,cm^{-1}\,K^{-1}}$, $h = 1.3 \times 10^{-3}\,\mathrm{cm}$, $R_0/R = 2.1$, $R_0 = R_2 = 0.1\,\mathrm{cm}$ and taking $R_2/R_1 = 2$) (ref. 40). Therefore, it is reasonable to expect that a much larger temperature difference will manifest itself across the droplet during vaporization than across the textured surface. With the resistance values, we estimate the temperature difference across the thin outer layer of the droplet to be $\Delta T \approx q\mathcal{R}_w$. Substituting appropriate values yields $\Delta T \approx -28\,\mathrm{K}$, so the estimated temperature difference is indeed substantial ($J = 0.69 \times 10^{-3}\,\mathrm{g\,cm^{-2}\,s^{-1}}$, $\Delta H_{vap} = 2{,}441\,\mathrm{J\,g^{-1}}$) (ref. 40).

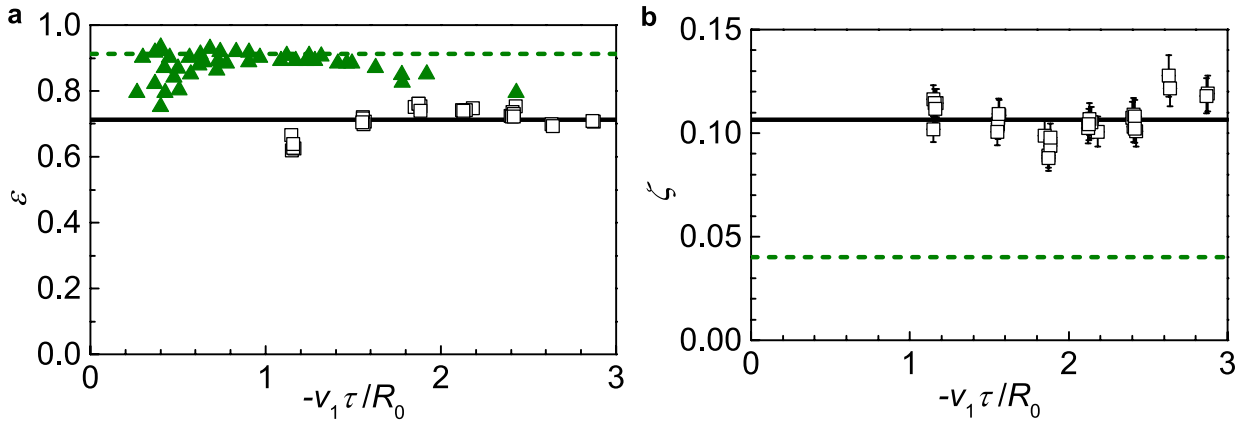
31. Maitra, T. *et al.* Hierarchically nanotextured surfaces maintaining superhydrophobicity under severely adverse conditions. *Nanoscale* **6,** 8710–8719 (2014).
32. Das, A., Schutzius, T. M., Bayer, I. S. & Megaridis, C. M. Superoleophobic and conductive carbon nanofiber/fluoropolymer composite films. *Carbon* **50,** 1346–1354 (2012).
33. Okumura, K., Chevy, F., Richard, D., Quéré, D. & Clanet, C. Water spring: a model for bouncing drops. *Europhys. Lett.* **62,** 237–243 (2003).
34. Chevy, F., Chepelianskii, A., Quéré, D. & Raphaël, E. Liquid Hertz contact: softness of weakly deformed drops on non-wetting substrates. *Europhys. Lett.* **100,** 54002 (2012).
35. Morse, D. C. & Witten, T. A. Droplet elasticity in weakly compressed emulsions. *Europhys. Lett.* **22,** 549–555 (1993).
36. Moláček, J. & Bush, J. W. M. A quasi-static model of drop impact. *Phys. Fluids* **24,** 127103 (2012).
37. Clanet, C., Béguin, C., Richard, D. & Quéré, D. Maximal deformation of an impacting drop. *J. Fluid Mech.* **517,** 199–208 (2004).
38. Poling, B. E., Prausnitz, J. M. & O'Connell, J. P. *The Properties of Gases and Liquids* Appendix B (McGaw-Hill, 2001).
39. Maxwell, J. C. *The Scientific Papers of James Clerk Maxwell* (ed. Niven, W. D.) Vol. 2 681–712 (Dover Publications, New York, 1890).
40. Lide, D. R. (ed.) *CRC Handbook of Chemistry and Physics* 85th edn (CRC Press, 2005).



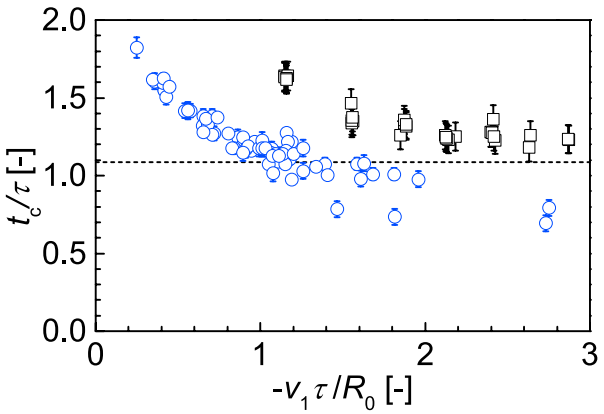
Extended Data Figure 1 | Schematic idealizing the droplet trampolining phenomenon as a hybrid MSD–projectile system. MSD and projectile motion apply when $y < 0$ and $y \geq 0$, respectively. The variables are mass m, droplet 'stiffness' k, damping coefficient c, initial droplet velocity v_0, droplet impact velocity v_1 and droplet recoil velocity v_2; $f(t)$ is the *f*orcing function. The horizontal dashed line indicates where y is zero.



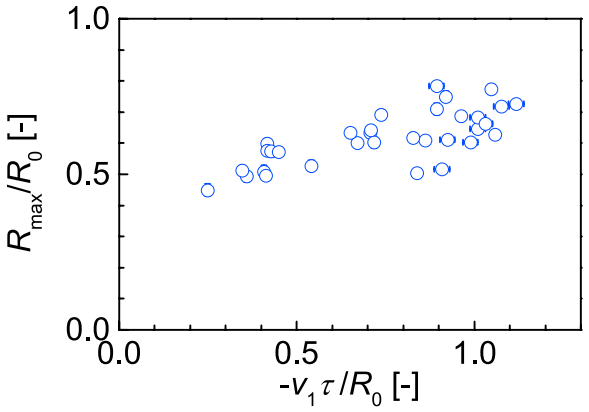
Extended Data Figure 2 | Comparing experimental and theoretical results for droplet trampolining. Quantities plotted are dimensionless. **a**, Plot of y as a function of t for experimental (blue circles) and theoretical (black line) cases. Inset, the droplet at the moment of impact (note that it is non-spherical). **b**, The applied force f required to generate the theoretical solutions in **a** as a function of time t. The magnitude of this force $\overline{f}$ was determined iteratively by matching the value of ε from the theory with that from the corresponding experiments. The impact properties for the droplet shown in **a** are $\mathrm{Bo} = mg/\sigma R_0 = 0.42$ and $v_1 = -0.5R_0/\tau$ (first impact). The properties of the superhydrophobic surface were $[d, l, h] = [1.4, 6.5, 4.8]$ μm.



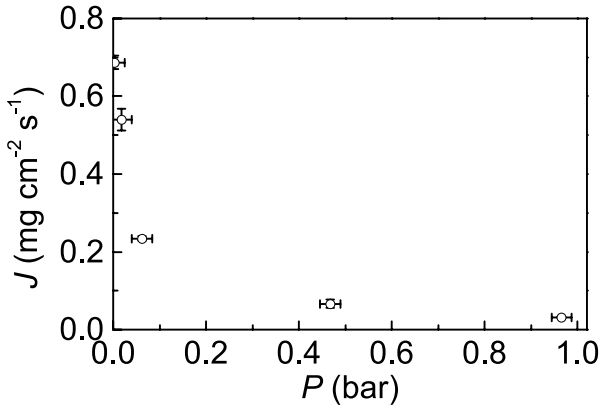
Extended Data Figure 3 | Determining the damping ratio ζ for droplets impacting superhydrophobic surfaces under standard pressure conditions. a, **b**, Plots of $-v_1\tau/R_0$ versus ε (**a**) and ζ (**b**) as determined from experiments on superhydrophobic surfaces. Square symbols represent experiments performed in this work (advancing and receding contact angles $\theta_a^* = 161^\circ \pm 3^\circ$, $\theta_r^* = 150^\circ \pm 4^\circ$; $[d, l, h] = [1.5, 6.5, 13.3]$ μm); errors represent the standard deviation of the measurement and triangles are experimental data from ref. 19 ($\theta_a^* \approx 170^\circ \pm 3^\circ$ and $\theta_a^* - \theta_r^* \approx 5^\circ$). In **a**, the dashed green line represents the theoretical upper limit for ε for droplet impact ($\sqrt{5/6}$); the solid black line is the average value of ε from the experiments performed in this study. In **b**, the solid black and dashed green lines represent the average values of ζ obtained from experiments in this study and the theoretical lower limit of ζ, respectively. The theoretical lower limit is estimated with using $\varepsilon = \sqrt{5/6}$ and $t_c/\tau = 1.09$ (ref. 19). Error bars in the plots represent measurement uncertainty.



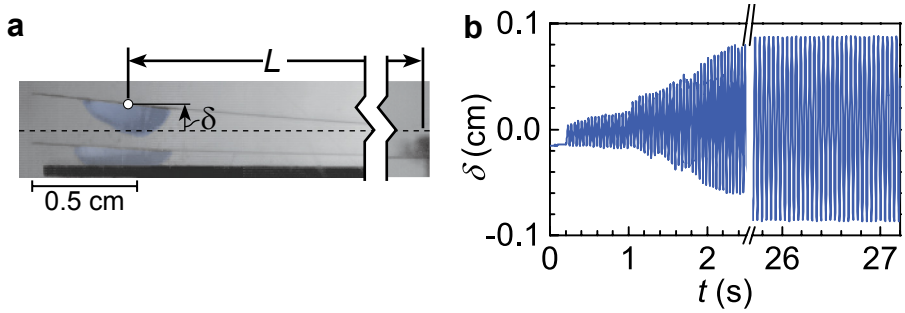
Extended Data Figure 4 | The role of environmental pressure on the contact time of a droplet with a superhydrophobic substrate for a single impact cycle. Plot of droplet–substrate contact time t_c/τ versus $-v_1\tau/R_0$ for water droplets impacting a superhydrophobic surface with a wetting fraction of $\phi = 0.04$ under low-pressure (circles) and standard-pressure (squares) conditions. The properties of the superhydrophobic surfaces were $[d, l, h] = [1.5, 6.5, 13.3]$ μm (squares) and $[d, l, h] = [1.4, 6.5, 4.8]$ μm (circles). The horizontal dashed line denotes the so-called minimum contact time $t_c/\tau \approx 1.09$. Error bars in the plots represent measurement uncertainty.



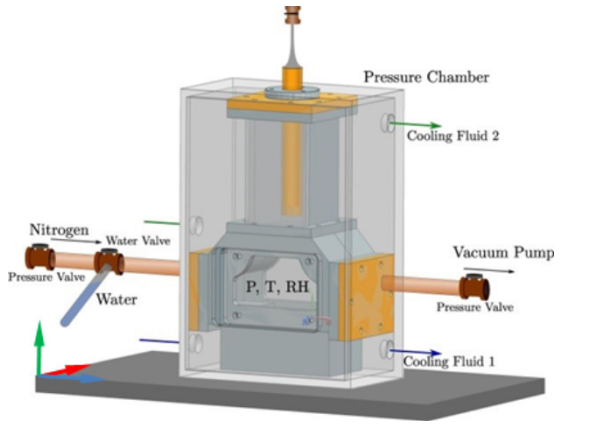
Extended Data Figure 5 | Spreading behaviour of a water droplet. Plot of R_{max}/R_0 versus $-v_1\tau/R_0$ for droplets impacting onto a superhydrophobic surface in a low-pressure, low-humidity environment. The properties of the superhydrophobic surface were $[d, l, h] = [1.4, 6.5, 4.8]$ μm. Error bars in the plots represent measurement uncertainty.



Extended Data Figure 6 | The role of environmental pressure on the vaporization flux of a water droplet in a low-humidity environment. Plot of vaporization flux J versus environmental pressure P for a millimetre-scale water droplet in contact with a superhydrophobic surface. The properties of the surface used were $[d, l, h] = [1.4, 6.5, 18.2]$ μm. Error bars for P and J represent the uncertainty of the measurement and s.d., respectively. Each data point is the average of five measurements.

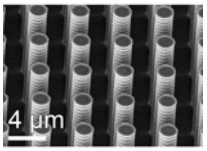
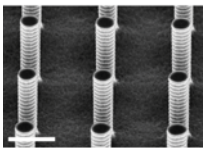
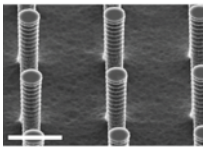
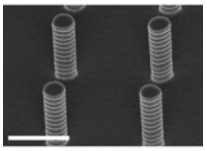
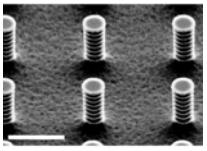
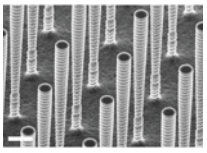
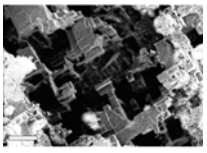
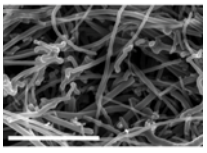


Extended Data Figure 7 | Exploiting trampolining dynamics with a cantilever. a, Overlaid image sequence (20 ms between the two images) of a droplet attached to a cantilever beam of length L exploiting droplet trampolining to create mechanical motion. **b**, Plot of beam deflection δ as a function of t for a similar sequence to that in **a**. See Supplementary Video 5 for further details. The properties of the surface used were $[d, l, h] = [1.5, 6.5, 13.3]$ μm.



Extended Data Figure 8 | Schematic showing the environmental chamber used throughout the study. We generated dry conditions in the chamber with nitrogen (N_2), and the pressure was reduced with a vacuum pump. The front and back of the chamber were equipped with transparent windows that were removable to facilitate placement of substrates and droplets. The coordinates XC, YC and ZC are denoted by blue, red and green, respectively.

Extended Data Table 1 | Experimental details on the engineered surfaces used in this study

Type	Processing technique	Specifications	ϕ	θ_a^* (°)	θ_r^* (°)	Micrograph
Silicon micropillar	Lithography and etching	$[d, l, h]$ = [2.0,4.6,13.5]μm	0.15	162±1	146±5	
Silicon micropillar	Lithography and etching	$[d, l, h]$ = [1.5,6.5,13.3] μm	0.04	161±3	150±4	
Silicon micropillar	Lithography and etching	$[d, l, h]$ = [1.6,6.5,10.9] μm	0.05	161±2	154±3	
Silicon micropillar	Lithography and etching	$[d, l, h]$ = [1.4,6.5,4.8] μm	0.04	161±2	154±2	
Silicon micropillar	Lithography and etching	$[d, l, h]$ = [1.6;6.5;3.5] μm	0.05	164±2	161±5	
Silicon micropillar	Lithography and etching	$[d, l, h]$ = [1.4;6.5;18.2] μm	0.04	166±1	162±1	
Etched aluminum	Etching	--	--	155±2	152±3	
Polymer nanocomposite	Spray coating	length of carbon nanofiber: 20-200 μm	--	151±2	148±2	

For micropillar surfaces, the pillar diameter, pitch and height are given by [*d*, *l*, *h*], respectively. The liquid–solid area fraction is ϕ. The apparent advancing and receding contact angles are θ_a^* and θ_r^*, respectively. All scale bars are 4 μm.

Extended Data Table 2 | Experimental probability of ice levitation as a function of droplet size on the CNF–PMC coating under dry, low-pressure conditions for an environment at room temperature

$R_{0,\min}$ [mm]	$R_{0,\max}$ [mm]	Number of trials [num]	Probability of ice levitation [-]
0.65	0.74	5	0.2
0.88	1.18	5	1.0
1.30	1.33	5	0.8
1.47	1.51	5	1.0
1.59	1.69	5	0.8

LETTER

doi:10.1038/nature15691

Rhodium-catalysed *syn*-carboamination of alkenes via a transient directing group

Tiffany Piou[1] & Tomislav Rovis[1]

Alkenes are the most ubiquitous prochiral functional groups—those that can be converted from achiral to chiral in a single step—that are accessible to synthetic chemists. For this reason, difunctionalization reactions of alkenes (whereby two functional groups are added to the same double bond) are particularly important, as they can be used to produce highly complex molecular architectures[1,2]. Stereoselective oxidation reactions, including dihydroxylation, aminohydroxylation and halogenation[3–6], are well established methods for functionalizing alkenes. However, the intermolecular incorporation of both carbon- and nitrogen-based functionalities stereoselectively across an alkene has not been reported. Here we describe the rhodium-catalysed carboamination of alkenes at the same (*syn*) face of a double bond, initiated by a carbon–hydrogen activation event that uses enoxyphthalimides as the source of both the carbon and the nitrogen functionalities. The reaction methodology allows for the intermolecular, stereospecific formation of one carbon–carbon and one carbon–nitrogen bond across an alkene, which is, to our knowledge, unprecedented. The reaction design involves the *in situ* generation of a bidentate directing group and the use of a new cyclopentadienyl ligand to control the reactivity of rhodium. The results provide a new way of synthesizing functionalized alkenes, and should lead to the convergent and stereoselective assembly of amine-containing acyclic molecules.

Functional groups that are based on nitrogen are prominent in biologically relevant molecules[7], and stereoselective chemical methods for introducing nitrogen atoms into organic molecules are the subject of intense interest. Alkene hydroamination—the addition of a nitrogen and hydrogen across a carbon–carbon double bond—is an emerging technology for introducing nitrogen functionality (Fig. 1a)[8–10]. However, the incorporation of carbon-based coupling partners is more limited, despite the crucial role of reactions that form carbon–carbon bonds in chemical synthesis. Among these, Heck-type approaches are noteworthy for their ability to introduce a carbon fragment in a stereoselective manner under typically mild conditions[11,12]. But both the hydroamination and the Heck-type reactions have the same strategic drawback: only one end of the alkene is functionalized. Simultaneous incorporation of both carbon- and nitrogen-based functionalities (carboamination) across an alkene would address this deficiency.

Established stereoselective carboamination reactions are limited and fall into three categories (Fig. 1b). Of these, annulative reactions are popular and powerful but deliver a cyclic product, which limits their impact[13,14]. A handful of intramolecular approaches have also been developed, wherein one of the reacting partners is tethered to the alkene[15–17]. Finally, there is a growing subset of radical-based reactions, which functionalize both ends of the alkene in a carboamination process[18,19]. However, the involvement of radicals means that the stereochemistry present in the alkene starting material is typically lost. Here, we describe the stereoselective intermolecular carboamination of alkenes, using enoxyphthalimides as the source of both the carbon and the nitrogen atoms (Fig. 1c). In the presence of an Rh(III) catalyst, these precursors undergo stereospecific *syn* addition (addition to the same face of a double bond) to a variety of disubstituted alkenes, delivering acyclic products containing two contiguous stereocentres in an intermolecular fashion.

We have previously shown that enoxyphthalimides undergo Rh(III)-catalysed reactions with electron-deficient alkenes to deliver cyclopropane adducts (Fig. 2)[20]. The mechanism proposed involves the generation of intermediate **A**, the product of carborhodation of the alkene partner. We hypothesized that the Rh atom is coordinatively unsaturated and thus ligates the enol alkene fragment, which subsequently undergoes migratory insertion to form the carbon–carbon bond in the cyclopropane product. Should the Rh atom instead be coordinatively saturated, intramolecular alkene coordination should be disfavoured, and reductive elimination to form the carboamination product might be favoured. Coordinative saturation of the Rh atom could conceivably occur by intramolecular coordination to a bidentate directing group.

Our past efforts to install requisite bidentate directing groups[21] on the enoxyamine were frustrated by the instability of the product. We overcame this instability by generating a bidentate directing group

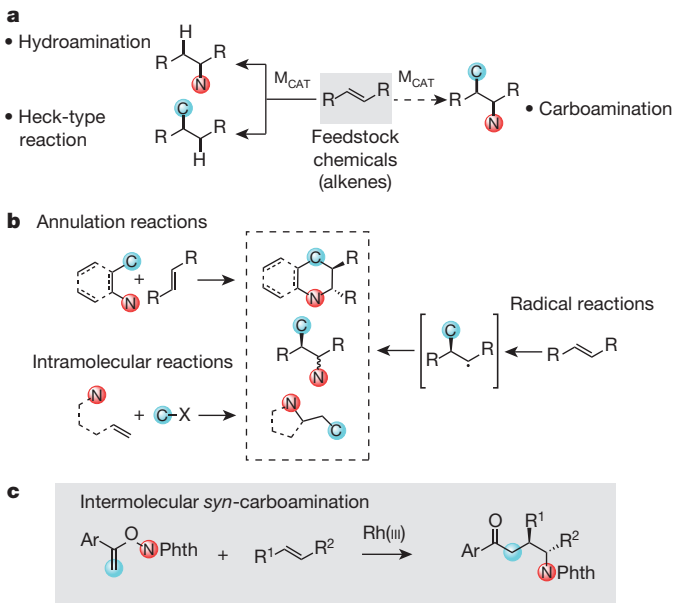


Figure 1 | Carboamination reactions. a, Transition-metal-catalysed difunctionalization of alkenes. Previously, such reactions could reliably achieve the introduction of either nitrogen-based or carbon-based functional groups (left-hand reaction); known reactions that introduce both groups across a single alkene (carboamination reactions, right, dotted arrow) have drawbacks. M_{CAT}, metal-based catalyst; R, functional group. **b**, The previously known carboamination reactions in organic synthesis: annulation reactions, intramolecular reactions and radical reactions, all of which have limitations. **c**, Our proposed Rh(III)-catalysed intermolecular *syn*-carboamination of alkenes. Ar, aromatic groups; Ph, phenyl; Phth, phthalimide.

[1]Department of Chemistry, Colorado State University, Fort Collins, Colorado 80523, USA.

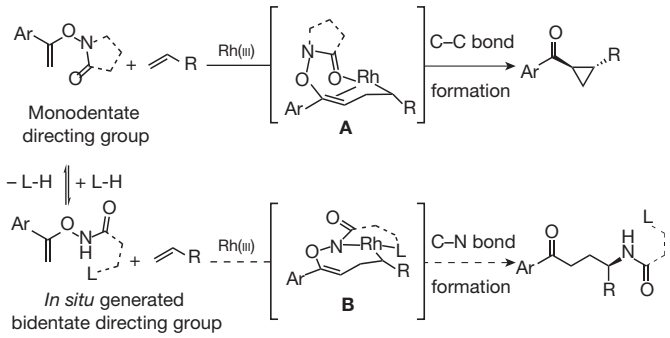


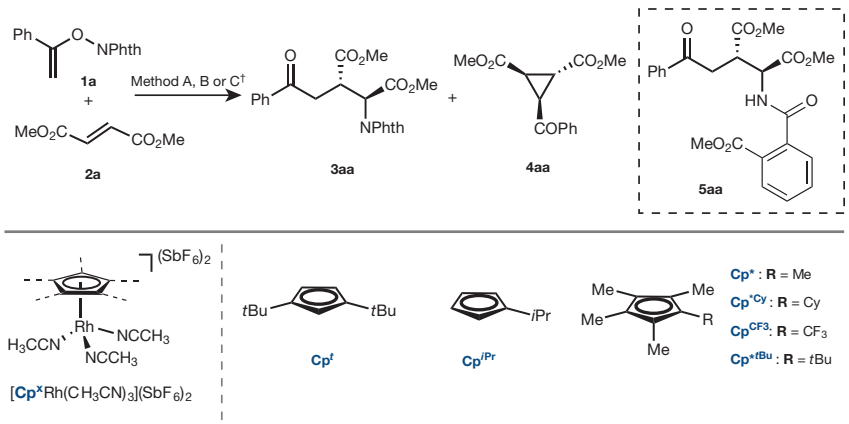
Figure 2 | Working hypothesis: tuning of the directing group to influence reactivity. The ligands on the Rh(III) catalyst are omitted for clarity. Ar, aromatic group; L, exogenous nucleophile. A monodentate directing group allows for intramolecular coordination of the alkene to Rh (intermediate **A**), leading to cyclopropanation. A bidentate directing group occupies the last coordination site on Rh (intermediate **B**), leading to reductive elimination to form a carboamination product.

in situ using a more nucleophilic solvent such as methanol, which we hypothesized would open the phthalimide to form the phthalimide-derived amido ester. Under these conditions, the formation of the carboamination product **3aa** is favoured over the cyclopropane **4aa** in a 2.8/1 ratio (Table 1, entry 2). We also observed the formation of the product **5aa**, derived from the opening of the phthalimide ring. Fortunately, the product **5aa** could be converted back to **3aa** without erosion of diastereoselectivity, simply by heating the crude reaction mixture at 60 °C in toluene after consumption of the starting material **1a** (entry 3). Furthermore, we established that **3aa** was formed as a single diastereoisomer, the relative configuration being unambiguously assigned by X-ray crystallography and consistent with a *syn*-addition process, thereby confirming our initial hypothesis. Selectivity between **3aa** and **4aa**, however, remained less than optimal.

Building on our previous work on cyclopentadienyl ligands, we speculated that control of the chemoselectivity could be achieved through ligand design. Disappointingly, however, when using the monosubtituted cyclopentadienyl isopropyl (Cp^{iPr}) ligand, which performed well in the cyclopropanation reaction, the carboamination product **3aa** is not formed (Table 1, entry 4). Sterically hindered (1,3-di-*tert*-cyclopentadienyl, Cp^{t}) (ref. 22) or electron-deficient (trifluoromethyl-tetra-methyl-cyclopentadienyl, Cp^{CF3}) (ref. 23) ligands furnish compound **3aa** in poor yields (entries 5 and 6). But the pentasubstituted ligand cyclohexyl-tetra-methyl-cyclopendienyl (Cp^{*Cy}) gives the desired product **3aa** with 69% yield and good chemoselectivity (**3aa**/**4aa** = 8.0/1; Table 1, entry 7). Further increasing the steric hindrance of the cyclopentadienyl ligand (*tert*-butyl-tetra-methyl-cyclopentadienyl, Cp^{*tBu}) allows the formation of **3aa** with an increased yield (72%) and slightly better chemoselectivity (**3aa**/**4aa** = 8.4/1; entry 8). Finally, replacing the base caesium acetate with caesium adamantylcarboxylate significantly improves the chemoselectivity (**3aa**/**4aa** = 14.8/1), producing the desired product **3aa** with an 82% yield (entry 9). Notably, decreasing the catalyst loading from 10 mol% to 5 mol% and using an equimolar amount of base did not affect the efficiency of the reaction (entry 10).

Having optimized the reaction conditions, we investigated the generality of the *syn*-carboamination (Fig. 3a). We first examined structural variations in the *N*-enoxyphthalimide (substrate **1**; Fig. 3b). The presence of a phenyl ring on substrate **1** proved essential. Electron-donating and electron-withdrawing substituents located at the *para*,

Table 1 | Optimization of reaction conditions



Entry	Method†	Cp^{x}	Solvent	Ratio **3aa**/**4aa**‡	Yield **3aa** (%)§
1	A	Cp*	Trifluoroethanol	1/2.3	30%‖
2	A	Cp*	Methanol	2.8/1	49%¶
3	B	Cp*	Methanol	3.5/1	60%
4	B	Cp^{iPr}	Methanol	—	0%
5	B	Cp^{t}	Methanol	—	<10%
6	B	Cp^{CF3}	Methanol	—	<10%
7	B	Cp^{*Cy}	Methanol	8.0/1	69%
8	B	Cp^{*tBu}	Methanol	8.4/1	72%
9	B#	Cp^{*tBu}	Methanol	14.8/1	82%
10	C	Cp^{*tBu}	Methanol	14.8/1	80%‖

†Method A: **1a** (1 equiv.), **2a** (1.2 equiv.), [Rh(III)] (10 mol%), caesium acetate (2 equiv.) in solvent (0.2 M), at room temperature for 16 hours. Method B: **1a** (1 equiv.), **2a** (1.2 equiv.), [Rh(III)] (10 mol%), caesium acetate (CsOAc; 2 equiv.) in solvent (0.2 M), at room temperature for 16 hours then stirred in toluene (0.2 M) at 60 °C for 4 hours. Method C: **1a** (1 equiv.), **2a** (1.2 equiv.), [Rh(III)] (5 mol%), 1-adamantyl carboxylate caesium (1-$AdCO_2Cs$; 1 equiv.) in methanol (0.2 M), at room temperature for 16 hours then stirred in toluene (0.2 M) at 60 °C for 4 hours. The reaction is shown at the top of the inset figure. The bottom-left part of the figure shows the prototypical structure of the Rh catalysts used here; the bottom-right part of the figure shows the defined structures of the cyclopentadienyl ligands.
‡Determined by analysis of the unpurified mixture by ^{1}H nuclear magnetic resonance (NMR) spectroscopy.
§NMR yield.
‖Isolated yield.
¶Ratio **3aa**/**5aa**/**4aa** = 2.8/1/1.
#1-$AdCO_2Cs$ was used as the base instead of CsOAc.
Cp, cyclopentadienyl; Cp*, penta-methyl-cyclopentadienyl; Cp^{CF3}, trifluoromethyl-tetra-methyl-cyclopentadienyl; Cp^{*Cy}, cyclohexyl-tetra-methyl-cyclopentadienyl; Cp^{iPr}, isopropyl cyclopentadienyl; Cp^{t}, 1,3-di-*tert*-butyl-cyclopentadienyl; Cp^{*tBu}, *tert*-butyl-tetra-methyl-cyclopentadienyl; Cy, cyclohexyl; *i*Pr, isopropyl; Me, methyl; Phth, phthalimide; Ph, phenyl; *t*Bu, *tert*-butyl.

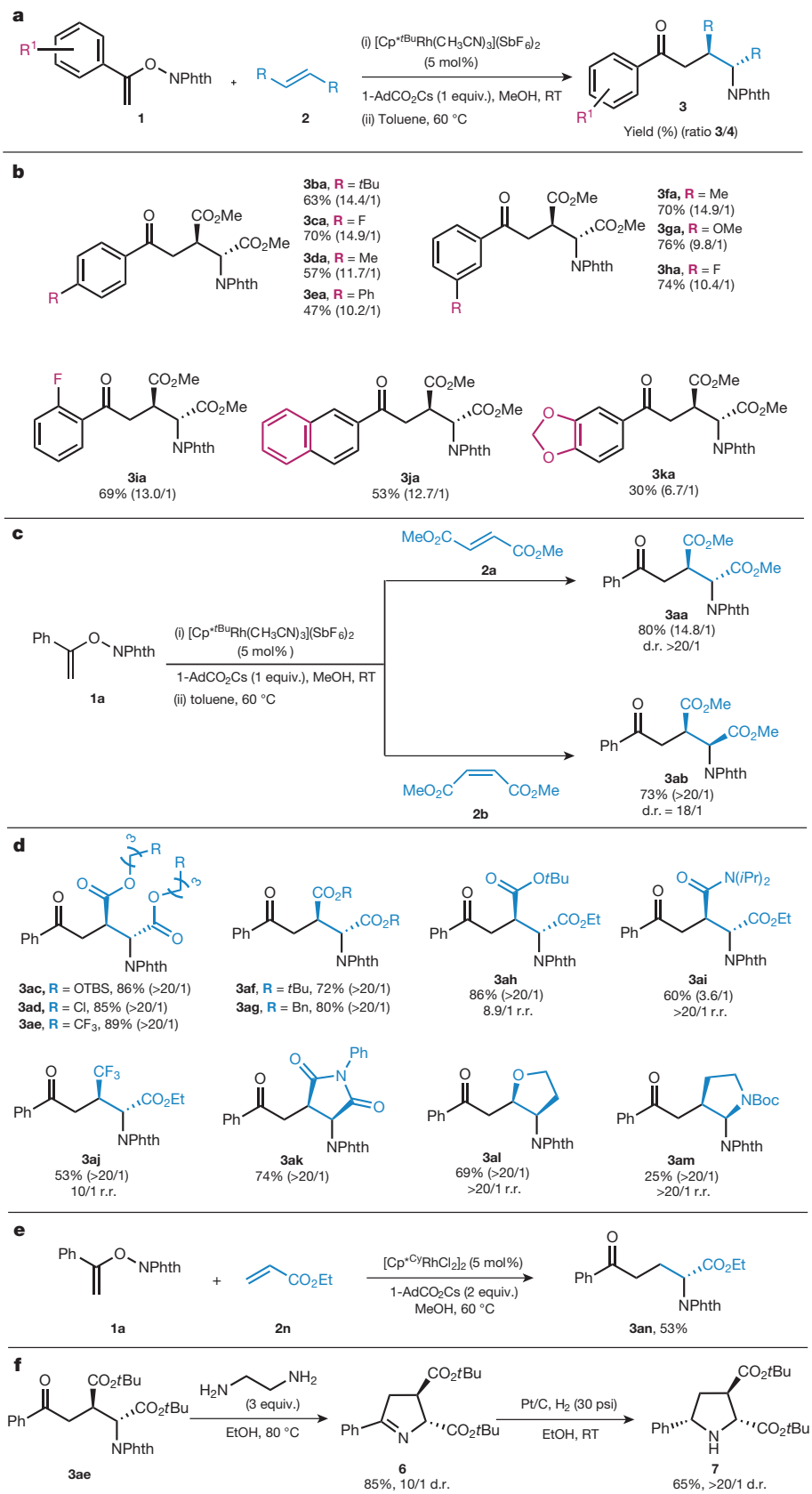


Figure 3 | Applications of the carboamination reaction. **a**, General conditions for carboamination of 1,2-disubstituted alkenes. **b**, Effect of substituents in the *N*-enoxyphthalimide. **c**, Probe of reaction stereospecificity. **d**, Functionalization of 1,2-disubstituted alkenes. **e**, Functionalization of monosubstituted alkene. **f**, Derivatization of the carboamination adduct: formation of pyrrolidine. Ac, acetyl; Ad, adamantyl; Bn, benzyl; Boc, *tert*-butoxycarbonyl; Cy, cyclohexyl; Cp, cyclopentadienyl; Cp*, penta-methyl-cyclopentadienyl; Et, ethyl; EtOH, ethanol; *i*Pr, isopropyl; Me, methyl; MeOH, methanol; Ph, phenyl; Phth, phthalimide; r. r., regioisomeric ratio; RT, room temperature; *t*Bu, *tert*-butyl; TBS, *tert*-butylsilyl. Yields are given as percentages; ratios of product **3**/product **4** are also given.

meta, or *ortho* positions of the phenyl ring are all well tolerated, providing the corresponding carboamination adducts **3ba**–**3ka** in good to high yields (30–76%). A few products were formed with low yields, which we think correlates with the relative insolubility of their derived starting materials (*N*-enoxyphthalimides **1e**, **1j** and **1k**).

In order to probe the stereochemical outcome of the reaction, we subjected fumarate and maleate esters (**2a** and **2b**, Fig. 3c) to the optimized reaction conditions. The reaction delivered isomeric products **3aa** and **3ab** in high diastereoselectivity, suggesting that the insertion event is a stereospecific *syn* addition across the alkene. We

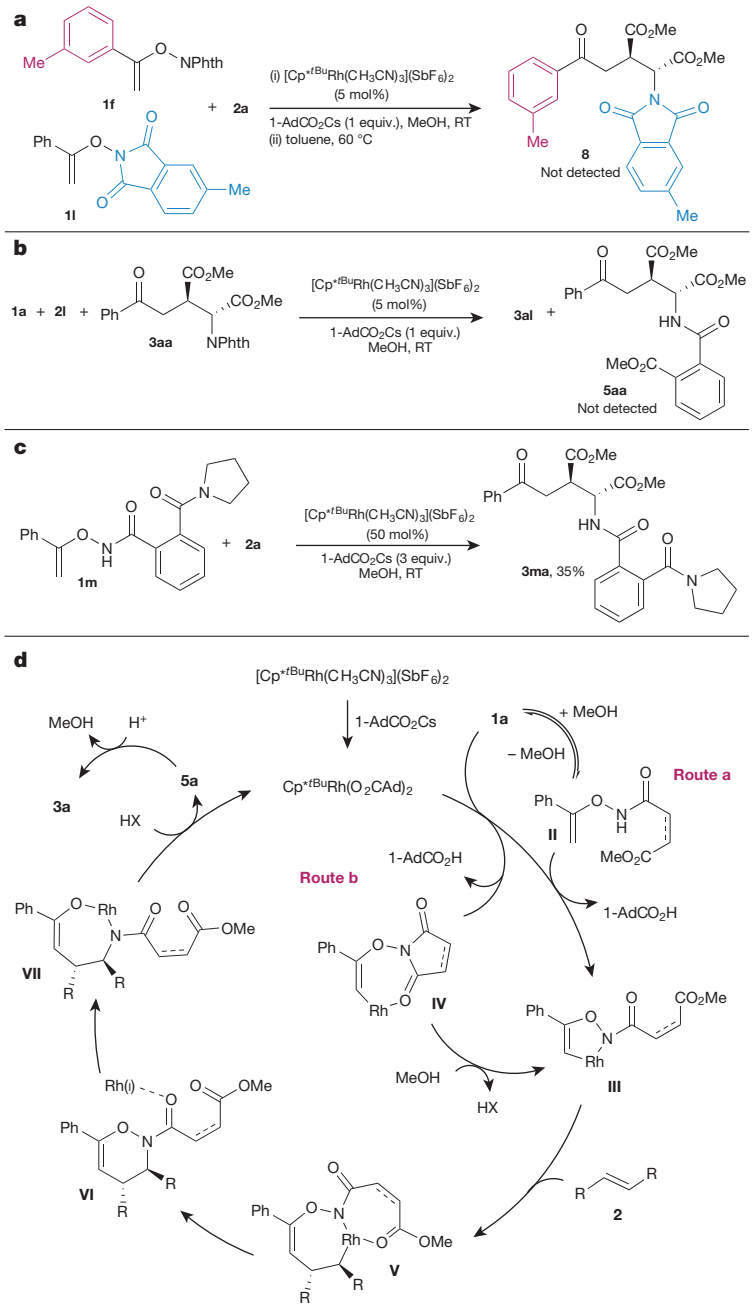


Figure 4 | Study and proposed reaction mechanism. **a**, Crossover experiment using an equimolar mixture of *N*-enoxyphthalimides **1f** and **1l**. No crossover adduct, **8**, is formed. **b**, An investigation of the formation of **5aa**: this does not form when product **3aa** is subjected back to the reaction. **c**, Reactivity of a secondary amide in the carboamination reaction; this finding suggests that the secondary enoxyamide is the active precursor to entering the catalytic cycle and supporting the potential intermediacy of intermediate **II** in **d**. **d**, Proposed mechanism for the carboamination reaction. In route a, in the presence of methanol (MeOH) and a base (1-$AdCO_2Cs$), the *N*-enoxyphthalimide **1a** can reversibly open to form intermediate **II**; the active Rh catalyst then undergoes an irreversible C–H bond activation to form intermediate **III**. (Alternatively, it is possible that the C–H activation precedes the opening of the phthalimide group, **IV** to **III**; route b.) Insertion of alkene **2** generates the coordinatively saturated intermediate **V**. Reductive elimination forms the C–N bond of intermediate **VI** while also reducing Rh(III) to Rh(I). Subsequent insertion of Rh(I) into the N–O bond delivers intermediate **VII**, which undergoes protonolysis to form **5a**, and turns the catalyst over. **5a** undergoes spontaneous slow cyclization to form **3a**. Ligands on Rh and phthalimide substituents are omitted for clarity. Ad, adamantyl; Cp, cyclopentadienyl, Me, methyl; Ph, phenyl; Phth, phthalimide; *t*Bu, *tert*-butyl; RT, room temperature.

next tested a variety of alkenes in our carboamination reaction (Fig. 3d); we found that the reaction conditions are mild enough to tolerate sensitive functional groups such as silyl ethers, chloro-alkyls and fluoro-alkyls. The corresponding adducts **3ac**–**3ae** were isolated in high yields (85–89%) with excellent chemoselectivity. The reaction also proceeds with hindered alkenes, leading to **3af** and **3ag** in excellent yields. Interestingly, in the case of unsymmetrical *trans*-1,2-disubstituted alkenes, the carboamination reaction takes place with a high control of regioselectivity, leading to products **3ah**–**3aj** as the major regioisomers (53–86% yield). In all cases, the most bulky substituent is placed away from the phthalimide group. Also, *N*-phenylmaleimide **2k** is a suitable substrate, giving the desired product **3ak** with a 74% yield. Electron-rich alkenes such as 1,2-dihydrofuran (**2l**) and 1,2-dihydropyrrole (**2m**) are also reactive, and produce disubstituted tetrahydrofuran (**3al**) and pyrrolidine (**3am**) with a 69% and a 25% yield, respectively. Gratifyingly, both heterocycles were obtained as single regioisomers and diastereoisomers, as found previously[24]. Finally, by switching to $[Cp^{*Cy}Rh(CH_3CN)_3](SbF_6)_2$ as the catalyst, the scope of the carboamination reaction was expanded to include monosubstituted alkenes. Thus, when using ethyl acrylate (**2n**) as a coupling partner, the unnatural α-aminoacid derivative **3an** is isolated with a 53% yield (Fig. 3e).

The carboamination products, **3**, are versatile entities. In addition to showing similarity to unnatural α-amino acids, they may also be converted into pyrrolidines (**7**; Fig. 3f). Deprotection of the phthalimide group followed by cyclization affords the 1,2-dihydropyrrole **6** (diastereomeric ratio, d.r. = 10/1, 85% yield), which can be reduced under heterogeneous conditions to yield pyrrolidine **7** in high diastereoselectivity (d.r. > 20/1).

In order to investigate the mechanism underlying the carboamination reaction, we probed whether delivery of the phthalimide moiety occurs through an intramolecular or intermolecular process. We carried out a crossover experiment by submitting an equimolar mixture of *N*-enoxyphthalimides **1f** and **1l** to our optimized reaction conditions (Fig. 4a). No crossover adduct **8** is formed, suggesting that, in agreement with our initial proposal (Fig. 2b), delivery of the phthalimide

moiety takes place intramolecularly. Moreover, when product **3aa** was subjected back to the reaction, product **5aa** did not form (Fig. 4b), suggesting that the phthalimide group opens before the final product **3aa** is formed. Thus, the adduct **5aa** might be formed first, and cyclizing back during the reaction to give **3aa**. We confirmed this assumption by monitoring the reaction progress using nuclear magnetic resonance (see Supplementary Information). To elaborate further on this idea, we investigated the reactivity of a bidentate substrate. Attempts to open the phthalimide group with methanol were unsuccessful owing to the instability of the product. However, the parent substrate **1m** proved more stable, and was subjected to the carboamination reaction (Fig. 4c). The expected product **3ma** was indeed formed, albeit with a moderate yield (35%). A control experiment demonstrates that the carboamination product **3ma** does not open in the presence of exogenous pyrrolidine under our standard reaction conditions (see Supplementary Information). Taken together, these results support our hypothesis that the directing group might be bidentate and emerge from *in situ* opening of the phthalimide moiety.

On the basis of these experiments, we propose the following catalytic cycle (Fig. 4d). First, in the presence of methanol and a base, the *N*-enoxyphthalimide **1a** can reversibly open to form intermediate **II** (Fig. 4d, route a). The active Rh(III) catalyst then undergoes an irreversible carbon–hydrogen activation at the alkene position, leading to the five-membered rhodacycle **III**. Alternatively, we cannot rule out the possibility that the carbon–hydrogen activation event precedes the opening of the phthalimide group (**IV** to **III**, Fig. 4d, route b). In either case, migratory insertion of alkene, **2**, then generates the coordinatively saturated Rh(III) complex **V**, with coordination of the ester group to the metal. We postulate that the bidentate directing group formed *in situ* stabilizes intermediate **V**, inhibiting both competitive migratory insertion into the enol alkene, and the β-H-elimination that forms the corresponding diene by a Heck-type process[25,26]. Instead, intermediate **V** undergoes reductive elimination to form intermediate **VI**. An oxidative addition of the nitrogen–oxygen bond into Rh(I) followed by protonation/tautomerization of the enol liberates the opened product **5a**, with concomitant regeneration of the active Rh(III) catalyst. Finally, during the reaction, the phthalimide group is re-formed to afford product **3a**. The origin of the chemoselectivity might be produced by the solvent effect (Table 1, entry 1 versus entry 2). When using methanol as the solvent, the initial opening of the phthalimide moiety prevails, favouring the formation of intermediate **III** and therefore the carboamination pathway. Conversely, the less-nucleophilic solvent trifluoroethanol tends to preserve the integrity of the phthalimide, and thus the cyclopropanation pathway is preferred.

We have developed a reaction that achieves *syn*-carboamination of disubstituted alkenes. The reaction uses enoxyphthalimides and a Rh(III) catalyst. Ligand development has revealed a new, bulky cyclopentadienyl group that alters the inherent chemoselectivity of a reaction. The use of methanol as a solvent is crucial, as is the observation that the phthalimide group undergoes *in situ* ring opening. Mechanistic experiments suggest that the basicity of the pendant carbonyl stabilizes a Rh(III) intermediate by coordinative saturation, leading to reductive elimination rather than to cyclopropanation. We are now investigating ways to broaden this reaction and to develop an asymmetric version of the transformation.

Received 2 June; accepted 4 September 2015.
Published online 21 October 2015.

1. McDonald, R. I., Liu, G. & Stahl, S. S. Palladium(II)-catalyzed alkene functionalization via nucleopalladation: stereochemical pathways and enantioselective catalytic applications. *Chem. Rev.* **111,** 2981–3019 (2011).
2. Chemler, S. R. & Bovino, M. T. Catalytic aminohalogenation of alkenes and alkynes. *Am. Chem. Soc. Catal.* **3,** 1076–1091 (2013).
3. Berkesell, A. & Gröger, H. *Asymmetric Organocatalysis* (Wiley-VCH, 2005).
4. Jacobsen, E. N. & Wu, M. H. *in Comprehensive Asymmetric Catalysis* (eds Jacobsen, E. N., Pfaltz, A. & Yamamoto, H.) 1309–1326 (Springer, 1999).
5. Hennecke, U. New catalytic approaches towards the enantioselective halogenation of alkenes. *Chem. Asian J.* **7,** 456–465 (2012).
6. Tan, C. K., Yu, W. Z. & Yeung, Y. Y. Stereoselective bromofunctionalization of alkenes. *Chirality* **26,** 328–343 (2014).
7. Vitaku, E., Smith, D. T. & Njardarson, J. T. Analysis of the structural diversity, substitution patterns, and frequency of nitrogen heterocycles among U.S. FDA approved pharmaceuticals. *J. Med. Chem.* **57,** 10257–10274 (2014).
8. Zhou, J. & Hartwig, J. F. Intermolecular, catalytic asymmetric hydroamination of bicyclic alkenes and dienes in high yield and enantioselectivity. *J. Am. Chem. Soc.* **130,** 12220–12221 (2008).
9. Shen, X. & Buchwald, S. L. Rhodium-catalyzed asymmetric intramolecular hydroamination of unactivated alkenes. *Angew. Chem. Int. Edn* **49,** 564–567 (2010).
10. Beller, M., Seayad, J., Tillack, A. & Jiao, H. Catalytic Markovnikov and anti-Markovnikov functionalization of alkenes and alkynes: recent developments and trends. *Angew. Chem. Int. Edn* **43,** 3368–3398 (2004).
11. Beletskaya, I. P. & Cheprakov, A. V. The Heck reaction as a sharpening stone of palladium catalysis. *Chem. Rev.* **100,** 3009–3066 (2000).
12. Werner, E. W., Mei, T.-S., Burckle, A. J. & Sigman, M. S. Enantioselective Heck arylations of acyclic alkenyl alcohols using a redox-relay strategy. *Science* **338,** 1455–1458 (2012).
13. Coldham, I. & Hufton, R. Intramolecular dipolar cycloaddition reactions of azomethine ylides. *Chem. Rev.* **105,** 2765–2810 (2005).
14. Nakamura, I. & Yamamoto, Y. Transition-metal-catalyzed reactions in heterocyclic synthesis. *Chem. Rev.* **104,** 2127–2198 (2004).
15. Mai, D. N. & Wolfe, J. P. Asymmetric palladium-catalyzed carboamination reactions for the synthesis of enantiomerically enriched 2-(arylmethyl)- and 2-(alkenylmethyl)pyrrolidines. *J. Am. Chem. Soc.* **132,** 12157–12159 (2010).
16. Wolfe, J. P. Synthesis of saturated heterocycles via metal-catalyzed alkene carboamination or carboalkoxylation reactions. *Top. Heterocycl. Chem.* **32,** 1–37 (2013).
17. Zeng, W. & Chemler, S. R. Copper(II)-catalyzed enantioselective intramolecular carboamination of alkenes. *J. Am. Chem. Soc.* **129,** 12948–12949 (2007).
18. Weidner, K., Giroult, A., Panchaud, P. & Renaud, P. Efficient carboazidation of alkenes using a radical desulfonylative azide transfer process. *J. Am. Chem. Soc.* **132,** 17511–17515 (2010).
19. Zhang, H. *et al.* Copper-catalyzed intermolecular aminocyanation and diamination of alkenes. *Angew. Chem. Int. Edn* **52,** 2529–2533 (2013).
20. Piou, T. & Rovis, T. Rh(III)-catalyzed cyclopropanation initiated by C–H activation: ligand development enables a diastereoselective [2 + 1] annulation of N-enoxyphthalimides and alkenes. *J. Am. Chem. Soc.* **136,** 11292–11295 (2014).
21. Mo, J., Wang, L., Liu, Y. & Cui, X. Transition-metal-catalyzed direct C–H functionalization under external-oxidant-free conditions. *Synthesis* 439–459 (2015).
22. Neely, J. M. & Rovis, T. Rh(III)-catalyzed regioselective synthesis of pyridines from alkenes and α,β-unsaturated oxime esters. *J. Am. Chem. Soc.* **135,** 66–69 (2013).
23. Hyster, T. K. & Rovis, T. An improved catalyst architecture for rhodium(III) catalyzed C–H activation and its application to pyridone synthesis. *Chem. Sci. (Camb.)* **2,** 1606–1610 (2011).
24. Webb, N. J., Marsden, S. P. & Raw, S. A. Rhodium(III)-catalyzed C–H activation/annulation with vinyl esters as an acetylene equivalent. *Org. Lett.* **16,** 4718–4721 (2014).
25. Guimond, N., Gorelsky, S. I. & Fagnou, K. Rhodium(III)-catalyzed heterocycle synthesis using an internal oxidant: improved reactivity and mechanistic studies. *J. Am. Chem. Soc.* **133,** 6449–6457 (2011).
26. Rakshit, S., Grohmann, C., Besset, T. & Glorius, F. Rh(III)-catalyzed directed C−H olefination using an oxidizing directing group: mild, efficient, and versatile. *J. Am. Chem. Soc.* **133,** 2350–2353 (2011).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank the National Institute of General Medical Sciences (grant no. GM80442) for support. We thank Johnson Matthey for rhodium salts, and J. Chu and B. Newell (at Colorado State University) for solving X-ray structures.

Author Contributions T.P. and T.R. conceived the concept and prepared the manuscript. T.R. directed the investigations. T.P. developed and studied the reaction.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to T.R. (rovis@colostate.edu).

LETTER

doi:10.1038/nature15380

The genetic sex-determination system predicts adult sex ratios in tetrapods

Ivett Pipoly[1], Veronika Bókony[1,2], Mark Kirkpatrick[3], Paul F. Donald[4,5], Tamás Székely[6]* & András Liker[1]*

The adult sex ratio (ASR) has critical effects on behaviour, ecology and population dynamics[1,2], but the causes of variation in ASRs are unclear[3,4]. Here we assess whether the type of genetic sex determination influences the ASR using data from 344 species in 117 families of tetrapods. We show that taxa with female heterogamety have a significantly more male-biased ASR (proportion of males: 0.55 ± 0.01 (mean ± s.e.m.)) than taxa with male heterogamety (0.43 ± 0.01). The genetic sex-determination system explains 24% of interspecific variation in ASRs in amphibians and 36% in reptiles. We consider several genetic factors that could contribute to this pattern, including meiotic drive and sex-linked deleterious mutations, but further work is needed to quantify their effects. Regardless of the mechanism, the effects of the genetic sex-determination system on the adult sex ratio are likely to have profound effects on the demography and social behaviour of tetrapods.

The adult sex ratio (ASR) varies widely in nature, ranging from populations that are heavily male-biased to those composed only of adult females[4–6]. Birds and schistosome parasites tend to have male-biased ASRs, for example, whereas mammals and copepods usually exhibit female-biased ASRs[4]. Extreme bias occurs among marsupials (Didelphidae and Dasyuridae): males die after the mating season, so there are times when the entire population consists of pregnant females[7]. Understanding the causes and consequences of ASR variation is an important goal in evolutionary biology, population demography and biodiversity conservation because the ASR affects behaviour, breeding systems and ultimately population fitness[1,2,8–10]. It is also an important issue in social sciences, human health and economics, since unbalanced ASRs have been linked to violence, rape, mate choice decisions and the spread of diseases such as HIV[11,12]. The causes of ASR variation in wild populations, however, remain obscure[4,8,13].

One factor that could affect the ASR is the genetic sex-determination system[5,6,14]. Taxa such as mammals and fruitflies (*Drosophila*) have XY sex determination (males are heterogametic), whereas taxa such as birds and butterflies have ZW sex determination (females are heterogametic). Sex-determination systems could affect the ASR in several ways. A skewed ASR might result from an unbalanced sex ratio at birth caused by sex ratio distorters[15]. Alternatively, a biased ASR could develop after birth if sex chromosomes contribute to sex differences in mortality[6,14,16]. Differential postnatal mortality is likely to be the main driver of biased ASRs in birds and mammals, since birth sex ratios in these classes tend to be balanced[5].

Here we use data from the four major clades of tetrapods (amphibians, reptiles, birds and mammals) to assess whether ASRs, measured by convention as the proportion of males in the population, differ between taxa with XY and ZW sex determination (Fig. 1 and Supplementary Data). While mammals and birds are fixed for XY and ZW sex determination, respectively, reptiles and amphibians provide particularly attractive opportunities for this study, since transitions between sex-determination systems have occurred many times within these clades[17]. We compiled published data on adult sex ratios in wild populations and their sex-determination systems (Supplementary Data). To control for phylogenetic effects, we used phylogenetic generalized least squares (PGLS)[18] models to test for differences in ASRs between XY and ZW taxa, and Pagel's discrete method (PDM)[19] to test whether XY and ZW systems are evolutionarily associated with female-biased and male-biased sex ratios, respectively. Phylogenies were taken from recent molecular studies (see Methods for details).

Both the ASR and the sex-determination system are highly variable across tetrapods (Fig. 1 and Supplementary Data). We find that the ASR and sex determination are correlated. Before controlling for phylogenetic effects, we find that ASRs are significantly more male-biased in species with ZW sex determination than in those with XY sex determination (Fig. 2, Table 1 and Extended Data Table 1). Similarly, the proportion of species with male-biased ASRs is greater among ZW than XY species (Fig. 1 and Table 1). These differences are significant within amphibians, within reptiles, and across tetrapods as a whole (Table 1 and Extended Data Table 1).

The pattern remains significant after controlling for phylogenetic effects. Both the mean of ASR across species (analysed using PGLS) and the proportion of species with male-biased sex ratios (analysed using PDM) differ significantly between XY and ZW systems within amphibians, within reptiles, and across tetrapods as a whole (Table 1 and Extended Data Table 1).The effect is strong in clades with variation in sex determination: the type of genetic sex determination explains up to 24% of the interspecific variance in the ASR among amphibians and 36% in reptiles (estimated using PGLS; Extended Data Table 2). The results remain significant when we treat three large clades with invariant sex-determination systems as a single datum each (snakes, ZW; birds, ZW; mammals, XY; Extended Data Table 1), when we make different assumptions about branch lengths in the phylogeny (Extended Data Table 2), and when we use arc-sine-transformed ASR values and control for variance in sample size (see Methods).

Body size and breeding latitude correlate with life-history traits in many organisms and these traits could affect ASR[20]. Sexual size dimorphism is linked to differential sexual selection acting on males and females and thus influences sex-specific mortality, and has been suggested to drive the evolution of genetic sex-determination systems[21]. Nevertheless, we find that neither body size nor breeding latitude explains significant variation in the ASR in phylogenetically controlled multi-predictor analyses (Table 2). Sexual size dimorphism is significantly associated with ASR in reptiles and across tetrapods as a whole, but the effect of the genetic sex-determination system remains significant when size dimorphism is included in the analysis (Table 2).

Sex differences in dispersal may also result in biased ASRs. However, dispersal is unlikely to explain the relationship between ASR and sex-determination systems. First, male-biased dispersal is typical in reptiles

[1]Department of Limnology, University of Pannonia, Pf. 158, H-8201 Veszprém, Hungary. [2]Lendület Evolutionary Ecology Research Group, Plant Protection Institute, Centre for Agricultural Research, Hungarian Academy of Sciences, Herman Ottó út 15, H-1022 Budapest, Hungary. [3]Department of Integrative Biology, University of Texas, Austin, Texas 78712, USA. [4]RSPB Centre for Conservation Science, RSPB, The Lodge, Sandy, Bedfordshire SG19 2DL, UK. [5]University of Cambridge, Conservation Science Group, Department of Zoology, Downing Street, Cambridge CB2 3EJ, UK. [6]Milner Centre for Evolution, Department of Biology and Biochemistry, University of Bath, Bath BA2 7AY, UK.
*These authors contributed equally to this work.

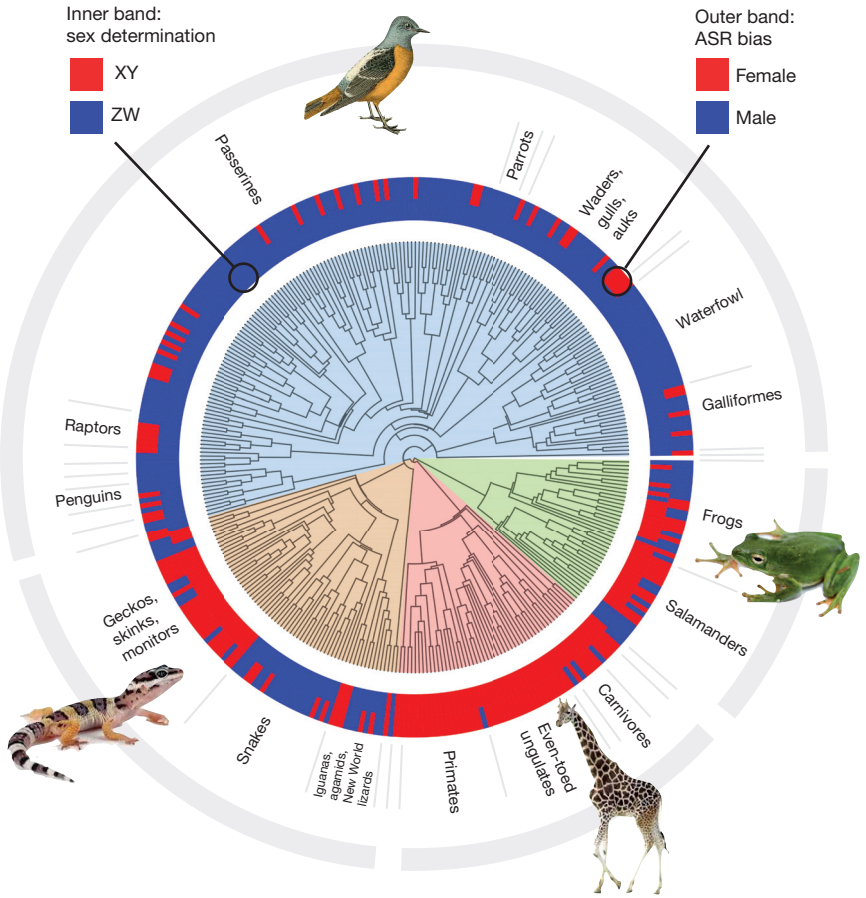


Figure 1 | Phylogenetic distribution of the ASR and genetic sex-determination systems across tetrapods. Inner band shows the type of sex determination (red: XY, blue: ZW), and the outer band shows the ASR bias for each species included in the study (red: ≤ 0.5, blue: > 0.5 proportion of males). Sample sizes: 39 species for amphibians, 67 species for reptiles, 187 species for birds and 51 species for mammals (see Supplementary Data).

regardless of the sex-determination system[22] (Supplementary Information 1). Second, there is no relationship between the ASR and sex bias in dispersal distance in birds (Supplementary Information 1). Finally, the relationship between sex determination and the ASR remains significant when the influence of sex-biased dispersal is controlled in multi-predictor models in tetrapods (Supplementary Information 1).

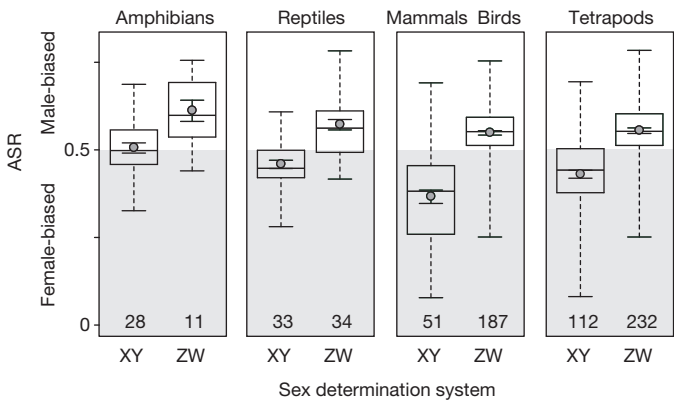


Figure 2 | Variation in the ASR as a function of the sex-determination system in amphibians, reptiles, mammals and birds, and across tetrapods (all four clades combined). Adult sex ratio is the proportion of males in all adults. Central dots and solid whiskers are mean ± s.e.m., horizontal bars are medians, and boxes and dashed whiskers show the interquartile ranges and data ranges, respectively, based on species values. Numbers of species are at the bottom of each panel. See Table 1 and Extended Data Table 1 for statistical results, and Extended Data Fig. 1 for phylogenetically corrected graphs.

The sex-determination system may affect the ASR in the directions seen in the data in several ways. First, sexual selection can fix mutations that increase male mating success and decrease male survival. These will accumulate on Y but not on W chromosomes, and will accumulate more readily on X than on Z chromosomes if they tend to be recessive. Second, biased ASRs could result from recessive mutations at loci carried on the X (or Z) chromosome but absent from the Y (or W) chromosome since they are not masked in the heterogametic sex (the 'unguarded sex chromosome' hypothesis)[5,6,14], and from deleterious mutations carried on the Y (or W) but not on the X (or Z) chromosome. At loci carried on both sex chromosomes, alleles on the Y (or W) can show partial degeneration[23]. Population genetic models suggest that deleterious mutation pressure alone may not be adequate to explain ASR biases as large as those observed (Supplementary Information 2), but the models do not include factors that could be important, notably the degeneration of Y and W chromosomes by genetic drift[23]. A third hypothesis is imperfect dosage compensation, which may be deleterious to the heterogametic sex[24]. Fourth, distorted sex ratios can result from meiotic drive acting on sex chromosomes[25]. Drive more often produces female-biased sex ratios in XY systems at birth[26]. There is little data on drive in ZW systems, but if it operates in a symmetrical fashion then we expect it to cause male-biased sex ratios. Fifth, the Y and W chromosomes might degenerate during the life-span, for example by telomere shortening or loss of epigenetic marks, more rapidly than the X and Z chromosomes. A final possibility is that sex-antagonistic selection acting on sex-linked loci could lead to biased

Table 1 | The effect of the sex-determination system on the ASR

Taxon	Number of species	Mean ASR				Species with male-biased ASR (%)		
		XY	ZW	*t*-test†	PGLS†	XY	ZW	PDM†
Amphibians	39	0.51	0.61	**	**	42.9	90.9	*
Reptiles	67	0.45	0.57	***	***	24.2	73.5	*
Birds	187	–	0.55	–	–	–	76.5	–
Mammals	51	0.37	–	–	–	9.8	–	–
Tetrapods	344	0.43	0.55	***	***	22.3	77.2	***

Mean ASR (proportion of males in the population), *t*-tests and the percentage of species with male-biased ASRs represent species-level statistics and analyses, while PGLS[18] and PDM[19] were used for phylogenetically corrected analyses of the difference in ASR between XY and ZW species.
* $P < 0.05$; ** $P < 0.01$; *** $P < 0.001$; '–' denotes no data or not tested.
† Detailed results of the statistical analyses are presented in Extended Data Table 1.

Table 2 | Phylogenetically corrected multi-predictor analyses of ASR variation

	Amphibians ($n = 39$)			Reptiles ($n = 67$)			Tetrapods ($n = 259$)		
	b (±s.e.m.)	*t*	*P*	*b* (±s.e.m.)	*t*	*P*	*b* (±s.e.m.)	*t*	*P*
Sex-determination system	0.10 (±0.03)	3.38	0.002	0.10 (±0.02)	4.56	<0.001	0.10 (±0.02)	5.23	<0.001
Body size	0 (±0)	1.41	0.166	0 (±0)	0.78	0.440	0 (±0)	0.05	0.962
Breeding latitude	0 (±0)	0.13	0.898	0 (±0)	0.04	0.966	0 (±0)	0.24	0.811
Sexual size dimorphism	−0.32 (±0.34)	0.92	0.363	−0.31 (±0.15)	2.17	0.034	−0.38 (±0.07)	5.57	<0.001

Relationships between the ASR, sex-determination system and other factors in phylogenetically corrected multi-predictor analyses using PGLS models[18]. Separate models of ASR were constructed for amphibians, reptiles and all tetrapods combined. For sex determination, *b* is the estimated difference in ASR between ZW and XY species.

sex ratios, but unlike the preceding hypotheses there does not seem to be a robust prediction about the direction of the ASR bias it will produce (Supplementary Information 2).

The limited data available do not provide clear support for any of these hypotheses, although critical tests are lacking. For instance, the meiotic drive process predicts biased sex ratios at birth. Although a recent comparative analysis in birds suggests that sex ratios at birth are unrelated to biased ASRs[8], offspring sex ratios have not been compared between different sex-determination systems. Further insight might come from the study of dioecious plants with biased sex ratios[27], but their skewed ASRs could result from selection on the gametophytic stage that is absent from animals[28]. Evolutionary feedbacks from the ASR to the sex-determination system are also possible: for example, the ASR could influence sexual size dimorphism and sexual conflict, which in turn could trigger transitions in sex determination[21,29,30].

In conclusion, we demonstrate strong and phylogenetically robust associations between genetic sex-determination systems and a demographic property of populations, the ASR. Although the mechanisms that drive this association need further theoretical and empirical analyses, the observed pattern is biologically important for two reasons. First, changes in sex-determination systems are expected to have knock-on effects on social behaviour. Theory suggests that the ASR affects violence, pair bonds, infidelity and parental care[1], and field-based studies support these predictions[3,10,12]. For instance, female-biased ASRs co-occur with polygyny and female care, whereas male-biased ASRs tend to co-occur with polyandry and male care in birds[3]. Second, sex-determination systems may have important demographic consequences through skewed birth sex ratios and sex-biased survival. Such biases may not only affect the productivity and growth of populations, but also their genetic composition and viability. Further theoretical, experimental and comparative studies are clearly needed to understand the linkages between sex determination, demography and social behaviour.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 19 February; accepted 5 August 2015.
Published online 7 October 2015.

1. Kokko, H. & Jennions, M. D. Parental investment, sexual selection and sex ratios. *J. Evol. Biol.* **21,** 919–948 (2008).
2. Le Galliard, J.-F., Fitze, P. S., Ferrière, R. & Clobert, J. Sex ratio bias, male aggression, and population collapse in lizards. *Proc. Natl Acad. Sci. USA* **102,** 18231–18236 (2005).
3. Liker, A., Freckleton, R. P. & Székely, T. The evolution of sex roles in birds is related to adult sex ratio. *Nature Commun.* **4,** 1587 (2013).
4. Székely, T., Weissing, F. J. & Komdeur, J. Adult sex ratio variation: implications for breeding system evolution. *J. Evol. Biol.* **27,** 1500–1512 (2014).
5. Donald, P. F. Adult sex ratios in wild bird populations. *Ibis* **149,** 671–692 (2007).
6. Trivers, R. L. in *Sexual Selection and the Descent of Men* (ed. Cambell, B.) 136–179 (Aldine, 1972).
7. Cockburn, A., Scott, M. P. & Dickman, C. R. Sex ratio and intrasexual kin competition in mammals. *Oecologia* **66,** 427–429 (1985).
8. Székely, T., Liker, A., Freckleton, R. P., Fichtel, C. & Kappeler, P. M. Sex-biased survival predicts adult sex ratio variation in wild birds. *Proc. R. Soc. B* **281,** 20140342 (2014).
9. Bessa-Gomes, C., Legendre, S. & Clobert, J. Allee effects, mating systems and the extinction risk in populations with two sexes. *Ecol. Lett.* **7,** 802–812 (2004).
10. Liker, A., Freckleton, R. P. & Székely, T. Divorce and infidelity are associated with skewed adult sex ratios in birds. *Curr. Biol.* **24,** 880–884 (2014).
11. Griskevicius, V. *et al.* The financial consequences of too many men: sex ratio effects on saving, borrowing, and spending. *J. Pers. Soc. Psychol.* **102,** 69–80 (2012).
12. Schacht, R., Rauch, K. L. & Borgerhoff Mulder, M. Too many men: the violence problem? *Trends Ecol. Evol.* **29,** 214–222 (2014).
13. Wilson, E. O. *Sociobiology: The New Synthesis* (Harvard Univ. Press, 1975).
14. Haldane, J. B. Sex-ratio and unisexual sterility in hybrid animals. *J. Genet.* **12,** 101–109 (1922).
15. Burt, A. & Trivers, R. *Genes in Conflict: The Biology of Selfish Genetic Elements* (Harvard Univ. Press, 2008).
16. Liker, A. & Székely, T. Mortality costs of sexual selection and parental care in natural populations of birds. *Evolution* **59,** 890–897 (2005).
17. Bachtrog, D. *et al.* Sex determination: why so many ways of doing it? *PLoS Biol.* **12,** e1001899 (2014).
18. Pagel, M. Inferring evolutionary processes from phylogenies. *Zool. Scr.* **26,** 331–348 (1997).
19. Pagel, M. Detecting correlated evolution on phylogenies: A general method for the comparative analysis of discrete characters. *Proc. R. Soc. Lond. B* **255,** 37–45 (1994).
20. Morrison, C. & Hero, J.-M. Geographic variation in life-history characteristics of amphibians: a review. *J. Anim. Ecol.* **72,** 270–279 (2003).
21. Adkins-Regan, E. & Reeve, H. K. Sexual dimorphism in body size and the origin of sex-determination systems. *Am. Nat.* **183,** 519–536 (2014).
22. Qi, Y., Yang, W., Lu, B. & Fu, J. Genetic evidence for male-biased dispersal in the Qinghai toad-headed agamid *Phrynocephalus vlangalii* and its potential link to individual social interactions. *Ecol. Evol.* **3,** 1219–1230 (2013).
23. Bachtrog, D. A dynamic view of sex chromosome evolution. *Curr. Opin. Genet. Dev.* **16,** 578–585 (2006).
24. Mank, J. E. Sex chromosome dosage compensation: definitely not for everyone. *Trends Genet.* **29,** 677–683 (2013).
25. Jaenike, J. Sex chromosome meiotic drive. *Annu. Rev. Ecol. Syst.* **32,** 25–49 (2001).
26. Werren, J. H. & Beukeboom, L. W. Sex determination, sex ratios, and genetic conflict. *Annu. Rev. Ecol. Evol. Syst.* **29,** 233–261 (1998).
27. Field, D. L., Pickup, M. & Barrett, S. C. H. Comparative analyses of sex-ratio variation in dioecious flowering plants. *Evolution* **67,** 661–672 (2013).

28. Hough, J., Immler, S., Barrett, S. C. H. & Otto, S. P. Evolutionarily stable sex ratios and mutation load. *Evolution* **67,** 1915–1925 (2013).
29. Roberts, R. B., Ser, J. R. & Kocher, T. D. Sexual conflict resolved by invasion of a novel sex determiner in Lake Malawi Cichlid fishes. *Science* **326,** 998–1001 (2009).
30. van Doorn, G. S. & Kirkpatrick, M. Transitions between male and female heterogamety caused by sex-antagonistic selection. *Genetics* **186,** 629–645 (2010).

Supplementary Information is available in the online version of the paper.

Acknowledgements M. Pennell and G. Imreh helped construct the phylogeny figure. We thank T. H. Clutton-Brock, S. P. Otto, D. Bachtrog and K. Reinhold for suggestions, and R. P. Freckleton for advice on analyses. We were supported by the European Union (TÁMOP-4.2.2.B-15/1/KONV-2015-0004), and by the US National Science Foundation (DEB-0819901 to M.K.). T.S. was supported by a Humboldt Award and MTA-DE 'Lendület' grant in projects that lead to the current work. A.L. was supported by the Hungarian Scientific Research Fund (OTKA K112838) and a Marie Curie Intra-European Fellowship.

Author Contributions T.S. and A.L. conceived the study. T.S., A.L. and V.B. designed the analyses. I.P., V.B., P.F.D. and A.L. collected the reptile, amphibian, mammalian and bird data, respectively. I.P., V.B. and A.L. conducted the analyses. M.K. developed the population genetic models. All authors wrote the paper.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to A.L. (aliker@almos.uni-pannon.hu).

METHODS

Data collection. We collected data on the ASR (expressed by convention as the proportion of males in the adult population) in amphibians and reptiles from literature published by December 2013, by searching in Google Scholar and Web of Science with the key words 'sex ratio' and 'reptile' or 'amphibian' or the scientific names of species. We also used reviews to identify additional data sources[31,32]. ASR data for mammals[5] were obtained from a similar search finished in 2007; and we used avian ASR estimates from our existing data set (supplementary information of ref. 10). We aimed to collect all ASR data that were available for amphibian and reptile species with known sex determination, so no statistical methods were used to pre-determine sample sizes. During the collection of ASR data for amphibians and reptiles, investigators were blinded to the type of sex determination. ASR data for birds and mammals were collected before the initiation of the current study, and for different purposes.

We specifically collected ASR data for amphibians and reptiles from studies that aimed to obtain representative estimates for the population composition and thus provide reliable sex ratio data[33]. These include either long-term demographic studies applying mark-recapture or culling methods (that is, each individual was counted only once) with similar capture probabilities for the sexes, or total population counts. When more than one measure was available, we used the total counts of individually marked animals over the study period because this may best approximate the overall ASR. We excluded studies in which the authors explicitly stated or speculated that their data may not represent the population-level ASR, or when the methods were not described in enough detail to assess the reliability of the ASR estimate. Moreover, we tested whether ASR estimates differed between sampling (hand-capture, trap, other) and marking (mark-recapture, culling) methods, and we found no such differences (linear mixed-effects model with species as random factor, sampling: $F_{(3,\ 105)} = 0.50$, $P = 0.683$; marking: $F_{(2,\ 105)} = 2.18$, $P = 0.118$; $n = 234$ records). When more than one estimate of the ASR was available for the same population (for example, from several yearly counts at the same location) we took their mean weighted by sample size. When more than one independent record was available for a species from different populations or studies, we used their simple mean. Weighted and non-weighted mean ASRs were highly correlated (amphibians: Pearson's $r = 0.973$, $P < 0.001$, $n = 35$ species; reptiles: $r = 0.995$, $P < 0.001$, $n = 60$ species); we used non-weighted averages because not all studies reported sample size.

We categorized the genetic sex-determination (GSD) systems of the species from published sources either as male-heterogametic (XY) or female-heterogametic (ZW). For amphibians, only species with known GSD systems were included[31,34], because GSD is an evolutionarily labile trait in amphibians; species within a genus or even populations within a species can differ in GSD system[35]. For reptiles, we included species for which the GSD was known either at the family level or at the species level if both XY and ZW systems were found in the family[34,36,37]. Our result for reptiles is not changed qualitatively by restricting our analyses to those species for which the GSD is known at species level[34], that is, when species for which we assumed the GSD based on other species in the family were excluded (difference between XY and ZW reptile species, PGLS model[18,38]: $b \pm$ s.e.m. $= 0.11 \pm 0.02$; $t = 4.70$, $P < 0.001$, $n = 26$; $R^2 = 0.479$). All birds were assigned to ZW, and all mammals to XY sex-determination systems[34].

We also collected data on three additional ecological and behavioural variables to control for their known correlation with the ASR and so reduce potential confounding effects in multi-predictor analyses. First, we used body size, which was measured as snout-to-vent length (in mm) for amphibians and squamates, and carapace length for the two turtle species, where possible from the same population for which ASR was reported. Head-body length was used for mammals ($n = 36$) (Encyclopedia of Life, http://www.eol.org). Since head-body length is not available for the vast majority of birds, we calculated this from the total body length by subtracting bill and tail length ($n = 133$; Supplementary Data). Where we had sex-specific data, the mean of male and female head-body length was used as body size variable in the analyses.

Second, we estimated sexual size dimorphism as $\log_{10}$(male body size) $-$ $\log_{10}$(female body size). For birds, we used body mass dimorphism (data available for $n = 181$ species)[39] owing to the lack of sex-specific body length data. The results of the multivariate PGLS model of tetrapods presented in Table 2 remain qualitatively the same when wing length dimorphism (data available for $n = 153$ species) is used for birds instead of body mass dimorphism (effect of sex determination: $b \pm$ s.e.m. $= -0.10 \pm 0.02$, $t = 4.97$, $P < 0.001$; body size: $b \pm$ s.e.m. $= 0 \pm 0$, $t = 0.06$, $P = 0.949$; latitude: $b \pm$ s.e.m. $= 0 \pm 0$, $t = 0.223$, $P = 0.823$; size dimorphism: $b \pm$ s.e.m. $= -0.52 \pm 0.12$, $t = 4.33$, $P < 0.001$; $n = 248$ species).

Third, we included breeding latitude[40,41] as the geographic coordinates of the ASR studies for amphibians and reptiles, taking absolute values to represent distance from the Equator in latitudinal degree. When the authors did not report latitude, we used Google Earth to estimate it on the basis of the description of the study site. For birds and mammals, we used the latitudinal midpoint of the breeding range of the species ($n = 182$ and 44 species, for birds and mammals, respectively; sources: V. Remes, A. Liker, R. Freckleton & T. Székely unpublished data for birds, and the PanTHERIA database for mammals[42], respectively). Mean values of these variables were used if multiple data of body size, latitude or size dimorphism per species were available.

Other possible confounding factors include the lifespan of individuals and sex-specific dispersal distances. First, longer average lifespan may lead to exaggeration of ASR bias. However, in species with available data[43], lifespan is unrelated to the ASR (PGLS, birds: $b \pm$ s.e.m. $= 0 \pm 0$, $t = 0.196$, $P = 0.845$, $n = 71$ species; mammals: $b \pm$ s.e.m. $= 0 \pm 0$, $t = 0.751$, $P = 0.457$, $n = 35$ species) and also to the absolute deviation of the ASR from 0.5 (that is, when assuming that longer lifespan can exaggerate ASR bias in either direction; birds: $b \pm$ s.e.m. $= 0 \pm 0$, $t = 1.543$, $P = 0.127$, $n = 71$ species; mammals: $b \pm$ s.e.m. $= 0 \pm 0$, $t = 0.180$, $P = 0.858$, $n = 35$ species). Second, sex-specific dispersal can bias the ASR owing to the higher mortality in the sex with longer dispersal distances. However, we found no evidence of a relationship of sex bias in dispersal either with the GSD in reptiles or with the ASR in birds (Supplementary Information 1). For these reasons, and because data on lifespan and/or sex-specific dispersal are not available for most species in our ASR data set, we did not include these variables in the main multi-predictor models.

Our final data set comprises data on 39 amphibian species and 67 reptile species (in total, $n = 229$ ASR records from different populations), 187 bird species and 51 mammalian species (a total of 344 species). We could not find body size and latitude data for some species, thus sample sizes were reduced in multi-predictor models. All species-level data and their sources are given in Supplementary Data.

Data analysis. To assess the reliability of the amphibian and reptile ASR estimates, we calculated the repeatability of ASR as the intraclass correlation coefficient (ICC) following ref. 44, using only those species for which we had at least two ASR estimates from different populations. These analyses show a moderate repeatability of ASR, and that a significant part of ASR variation is interspecific (amphibians: ICC $= 0.559$, $F_{(22,96)} = 7.27$, $P < 0.001$, $n = 23$ species, $n = 120$ records; reptiles: ICC $= 0.524$, $F_{(13,26)} = 4.11$, $P = 0.001$, $n = 14$ species, $n = 40$ records). For birds, our earlier analyses showed that 44% of the ASR variation was interspecific, and that the direction of ASR (that is, male- or female-biased) was highly conserved: in 44 out 55 species (80%), the direction of the ASR bias was the same for all repeated estimates[4]. For mammals, we did not find enough multiple ASR data within species to estimate repeatability.

In the comparative analyses we used the topology of ref. 45 for amphibians, a composite phylogeny for reptiles[46–48], ref. 49 for birds[10], the family-level relationships of ref. 50 and the genus/species level relationships of ref. 51 for mammals. For analyses across tetrapods, the branching topology between these four major clades was based on recent tetrapod phylogenies[52,53] (Fig. 1). Because we did not have branch length information for these composite phylogenies, we ran the analyses using arbitrary gradual branch lengths according to Nee's method[54]. However, our results remained consistent when we repeated the analyses with other branch length assumptions (Pagel's method and unit branch lengths[54]; Extended Data Table 2).

To test the association between ASR bias (male- versus female-biased) and GSD (XY versus ZW) in phylogenetically corrected analyses, we used PDM[19] as implemented in BayesTrait[55]. We used maximum likelihood methods to fit independent and dependent models for transitions in ASR bias and GSD states, and compared the fit of these two models by a likelihood ratio test[19]. To test the ASR difference between XY and ZW species, we used PGLS models with maximum likelihood estimates of Pagel's λ values[18] using the R[56] package 'caper'[38,57]. ASR was the response variable in all models, and the genetic sex-determination system was fitted as the predictor (Table 1 and Extended Data Table 1). The parameter estimate b shows the difference in ASR (proportion of males in the population) between ZW and XY species. To test the robustness of the bivariate results, we added body size, breeding latitude and sexual size dimorphism as predictors in multi-predictor models to control for their potential confounding effects (Table 2). As in earlier ASR studies[4,5], the distribution of ASR values did not deviate significantly from normal in the four clades separately as well as in tetrapods as a whole; our results remain qualitatively identical when ASR is arc-sine-transformed before PGLS analyses (amphibians: $b \pm$ s.e.m. $= 0.10 \pm 0.03$, $t_{37} = 3.44$, $P = 0.001$, $n = 39$; reptiles: $b \pm$ s.e.m. $= 0.12 \pm 0.02$, $t_{65} = 5.95$, $P < 0.001$, $n = 67$; tetrapods: $b \pm$ s.e.m. $= 0.11 \pm 0.02$, $t_{342} = 5.24$, $P < 0.001$, $n = 344$).

The difference between XY and ZW systems for tetrapods is not sensitive to the inclusion of large clades with uniform sex-determination systems (snakes and birds are all ZW, mammals are all XY) because it remains unchanged when each of these clades is reduced to a single datum of its mean ASR (PGLS: $b \pm$ s.e.m. $= 0.10 \pm 0.02$, $t = 5.07$, $P < 0.001$, $R^2 = 0.232$, $n = 87$). Furthermore, our result is also robust to between-species differences in sample size: when we added log(number of individuals) to the previous model, the effect of sex determination

remained significant ($b \pm$ s.e.m. $= 0.15 \pm 0.07$, $t = 2.08$, $P = 0.041$), while sample size had no significant effect on ASR ($b \pm$ s.e.m. $= 0 \pm 0.01$, $t = 0.35$, $P = 0.72$, $n = 78$). Furthermore, sample size was not a significant predictor of ASR when we added it as a fourth confounding variable in the full PGLS model ($b \pm$ s.e.m. $= 0 \pm 0.01$, $t = 1.16$, $P = 0.250$, $n = 78$), and the effect of other predictors remained qualitatively the same as in Table 2. Finally, the results do not change when we only used the most reliable ASR data (based on mark-recapture or culling methods): sex-determination system is significantly related to ASR in amphibians, reptiles tetrapods (PGLS results, amphibians: $b \pm$ s.e.m. $= 0.09 \pm 0.03$, $t = 3.07$, $P = 0.004$, $n = 35$ species; reptiles: $b \pm$ s.e.m. $= 0.11 \pm 0.03$, $t = 3.974$, $P < 0.001$, $n = 22$; tetrapods with snakes, birds and mammals included as single data points: $b \pm$ s.e.m. $= 0.10 \pm 0.02$, $t = 4.23$, $P < 0.001$, $n = 55$).

Population genetic models. We developed population genetic models of the effects that deleterious mutation and sex-antagonistic selection might have on the ASR (Supplementary Information 2). The models assume that deleterious mutations are largely or entirely recessive, that they have multiplicative fitness effects across loci, that the loci are fully sex-linked and in linkage equilibrium, that mutation is not sex-biased, and that selection is strong relative to mutation and drift. Fitness effects of mutations in hemizygotes and homozygotes are assumed equal. Full details of the models are given in Supplementary Information 2. Here we summarize the key results.

When deleterious alleles reach a mutation-selection balance, with XY sex determination the mean viability of males relative to females is

$$\overline{W}_{\mathrm{m}} \approx \exp\{-3U_{\mathrm{X}} - U_{\mathrm{Y}}\},$$

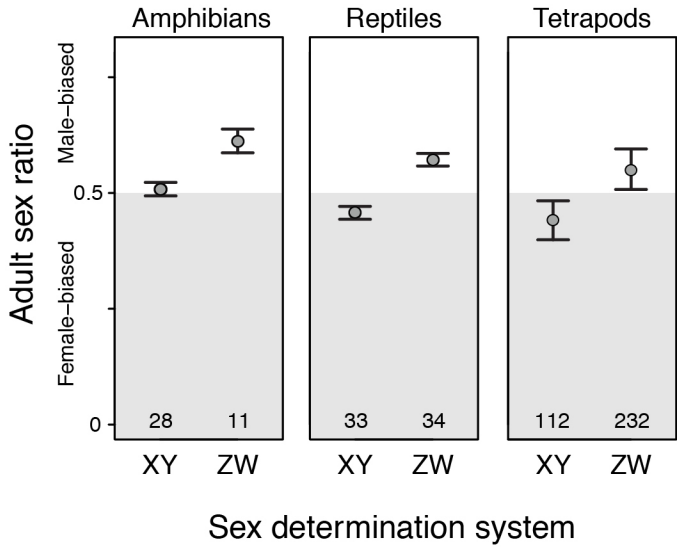
where U_{X} and U_{Y} are the total rates of mutation to deleterious alleles across all loci on the X and Y chromosomes. With ZW sex determination, the mean viability of females to males is

$$\overline{W}_{\mathrm{f}} \approx \exp\{-3U_{\mathrm{Z}} - U_{\mathrm{W}}\},$$

where U_{Z} and U_{W} are the total rates of mutation to deleterious alleles across all loci on the Z and W chromosomes. Using very rough estimates for rates of deleterious mutations appropriate for human sex chromosomes, we estimate that mutation-selection balance might bias the ASR by a few per cent. This degree of bias is substantially less than that seen in our data. We emphasize that the conclusion could be quite different using other parameter values, or if the model was extended to include stochastic effects.

The second hypothesis to explain biased ASRs that we explored with models is sex-antagonistic selection, the situation in which alleles are selected differently in females and males. In Supplementary Information 2, we use numerical examples to show that under both XY and ZW sex determination, either a female-biased or male-biased ASR can result. Thus there does not seem to be a robust generalization about how sex-antagonistic selection will bias the ASR.

31. Evans, B. J., Pyron, R. A. & Wiens, J. J. in *Polyploidy and Genome Evolution* (eds Soltis, P. S. & Soltis, D. E.) 385–410 (Springer Berlin Heidelberg, 2012).
32. Jongepier, E. *Reptilian Adult Sex Ratios are Biased Towards the Homogametic Sex.* Masters thesis, Univ. Groningen (2011).
33. Arendt, J. D., Reznick, D. N. & López-Sepulcre, A. Replicated origin of female-biased adult sex ratio in introduced populations of the Trinidadian Guppy (*Poecilia reticulata*). *Evolution* **68,** 2343–2356 (2014).
34. The Tree of Sex Consortium. Tree of Sex: A database of sexual systems. *Sci. Data* **1,** 140015 (2014).
35. Miura, I., Ohtani, H. & Ogata, M. Independent degeneration of W and Y sex chromosomes in frog *Rana rugosa*. *Chromosome Res.* **20,** 47–55 (2012).
36. Sarre, S. D., Ezaz, T. & Georges, A. Transitions between sex-determining systems in reptiles and amphibians. *Annu. Rev. Genomics Hum. Genet.* **12,** 391–406 (2011).
37. Pokorná, M. & Kratochvíl, L. Phylogeny of sex-determining mechanisms in squamate reptiles: are sex chromosomes an evolutionary trap? *Zool. J. Linn. Soc.* **156,** 168–183 (2009).
38. Freckleton, R. P., Harvey, P. H. & Pagel, M. Phylogenetic analysis and comparative data: a test and review of evidence. *Am. Nat.* **160,** 712–726 (2002).
39. Lislevand, T., Figuerola, J. & Székely, T. Avian body sizes in relation to fecundity, mating system, display behaviour and resource sharing. *Ecology* **88,** 1605 (2007).
40. Du, W., Robbins, T. R., Warner, D. A., Langkilde, T. & Shine, R. Latitudinal and seasonal variation in reproductive effort of the eastern fence lizard (*Sceloporus undulatus*). *Integr. Zool.* **9,** 360–371 (2014).
41. Iverson, J. B., Balgooyen, C. P., Byrd, K. K. & Lyddan, K. K. Latitudinal variation in egg and clutch size in turtles. *Can. J. Zool.* **71,** 2448–2461 (1993).
42. Jones, K. E. *et al.* PanTHERIA: a species-level database of life history, ecology, and geography of extant and recently extinct mammals. *Ecology* **90,** 2648 (2009).
43. Healy, K. *et al.* Ecology and mode of life explain lifespan variation in birds and mammals. *Proc. R. Soc. Lond. B* **281,** 20140298 (2014).
44. Lessells, C. M. & Boag, P. T. Unrepeatable repeatabilities? A common mistake. *Auk* **104,** 116–121 (1987).
45. Pyron, R. A. & Wiens, J. J. A large-scale phylogeny of Amphibia including over 2800 species, and a revised classification of extant frogs, salamanders, and caecilians. *Mol. Phylogenet. Evol.* **61,** 543–583 (2011).
46. Gardner, M. G., Hugall, A. F., Donnellan, S. C., Hutchinson, M. N. & Foster, R. Molecular systematics of social skinks: phylogeny and taxonomy of the *Egernia* group (Reptilia: Scincidae). *Zool. J. Linn. Soc.* **154,** 781–794 (2008).
47. Pyron, R. A., Burbrink, F. T. & Wiens, J. J. A phylogeny and revised classification of Squamata, including 4161 species of lizards and snakes. *BMC Evol. Biol.* **13,** 93 (2013).
48. Guillon, J.-M., Guery, L., Hulin, V. & Girondot, M. A large phylogeny of turtles (Testudines) using molecular data. *Contrib. Zool.* **81,** 147–158 (2012).
49. Jetz, W., Thomas, G. H., Joy, J. B., Hartmann, K. & Mooers, A. O. The global diversity of birds in space and time. *Nature* **491,** 444–448 (2012).
50. Meredith, R. W. *et al.* Impacts of the cretaceous terrestrial revolution and KPg extinction on mammal diversification. *Science* **334,** 521–524 (2011).
51. Fritz, S. A., Bininda-Emonds, O. R. P. & Purvis, A. Geographical variation in predictors of mammalian extinction risk: big is bad, but only in the tropics. *Ecol. Lett.* **12,** 538–549 (2009).
52. Chiari, Y., Cahais, V., Galtier, N. & Delsuc, F. Phylogenomic analyses support the position of turtles as the sister group of birds and crocodiles (Archosauria). *BMC Biol.* **10,** 65 (2012).
53. Amemiya, C. T. *et al.* The African coelacanth genome provides insights into tetrapod evolution. *Nature* **496,** 311–316 (2013).
54. Maddison, W. P. & Maddison, D. R. *Mesquite: a Modular System for Evolutionary Analysis*. http://mesquiteproject.org (2011).
55. Pagel, M. & Meade, A. Bayesian analysis of correlated evolution of discrete characters by reversible-jump Markov chain Monte Carlo. *Am. Nat.* **167,** 808–825 (2006).
56. R Development Core Team. *R: A Language and Environment for Statistical Computing* http://www.R-project.org (R Foundation for Statistical Computing, 2008).
57. Orme, A. D. *et al. caper: Comparative Analyses of Phylogenetics and Evolution in R* (v.0.5) https://cran.r-project.org/web/packages/caper/index.html (2013).



Extended Data Figure 1 | Phylogenetically corrected mean and s.e.m. of ASR in clades with different sex-determination systems. Parameter estimates for the mean and associated s.e.m. were calculated by PGLS models[18] presented in Extended Data Table 2 (with branch lengths estimated by Nee's method[54]).

Extended Data Table 1 | Detailed analyses of the effect of sex-determination system on the ASR.

	Species level		Phylogenetically corrected			
	t-test		PDM		PGLS	
	t-value	p-value (n)	LR	p-value (n)	t-value	p-value (n)
Amphibians (XY vs. ZW)	3.039	0.008 (39)	10.5	0.033 (39)	3.418	0.002 (39)
Reptiles (XY vs. ZW)	6.018	< 0.001 (67)	11.3	0.023 (67)	5.996	< 0.001 (67)
Mammals (XY) vs. birds (ZW)	8.982	< 0.001 (238)	not tested		not tested	
Tetrapods, all species (XY vs. ZW)	9.790	< 0.001 (344)	53.6	< 0.001 (344)	5.313	< 0.001 (344)
Tetrapods, reduced † (XY vs. ZW)	4.801	< 0.001 (87)	17.9	0.001 (87)	5.072	< 0.001 (87)

These are extensions of Table 1 showing details of the phylogenetically uncorrected (*t*-tests) and phylogenetically corrected (PGLS[18] and PDM[19]) analyses. Birds and mammals were not tested with phylogenetic control because there is no variation in the type of sex-determination system within birds and mammals. In the reduced analysis (marked by †), snakes, birds and mammals were each included as a single datum with mean species values.

Extended Data Table 2 | Phylogenetically controlled analyses of the relationship between ASR and genetic sex-determination system using different branch length assumptions.

Taxa	Branch lengths	b ± SE	t	p	R^2	λ
Amphibians	Nee's	0.101 ± 0.030	3.418	0.002	0.240	0.000
(n = 39)	Pagel's	0.101 ± 0.030	3.418	0.002	0.240	0.000
	Unit branch lengths	0.076 ± 0.027	2.821	0.008	0.177	0.000
Reptiles	Nee's	0.114 ± 0.019	5.996	< 0.001	0.356	0.000
(n = 67)	Pagel's	0.114 ± 0.019	5.968	< 0.001	0.354	0.000
	Unit branch lengths	0.114 ± 0.020	5.702	< 0.001	0.333	0.000
Tetrapods	Nee's	0.109 ± 0.020	5.313	< 0.001	0.076	0.409
(n = 344)	Pagel's	0.106 ± 0.021	4.998	< 0.001	0.068	0.332
	Unit branch lengths	0.093 ± 0.020	4.581	< 0.001	0.058	0.469

These are the results of PGLS models[18] as implemented in the R package 'caper'[57], showing parameter estimates (*b*) as the difference in ASR (ZW – XY), the proportion of interspecific variance (R^2) in ASR explained by the sex-determination system (female-heterogametic, ZW; or male-heterogametic, XY), calculated by PGLS; and the degree of phylogenetic dependence (λ). The models assume gradual branch lengths calculated either by Nee's or by Pagel's method, or unit branch lengths[54].

LETTER

doi:10.1038/nature15526

Differential responses to lithium in hyperexcitable neurons from patients with bipolar disorder

Jerome Mertens[1,2*], Qiu-Wen Wang[1*], Yongsung Kim[2], Diana X. Yu[2], Son Pham[2], Bo Yang[1], Yi Zheng[1], Kenneth E. Diffenderfer[3], Jian Zhang[4], Sheila Soltani[2], Tameji Eames[2], Simon T. Schafer[2], Leah Boyer[2], Maria C. Marchetto[2], John I. Nurnberger[5], Joseph R. Calabrese[6], Ketil J. Ødegaard[7], Michael J. McCarthy[8,9], Peter P. Zandi[10], Martin Alba[11], Caroline M. Nievergelt[9], The Pharmacogenomics of Bipolar Disorder Study†, Shuangli Mi[4], Kristen J. Brennand[12], John R. Kelsoe[8,9], Fred H. Gage[2] & Jun Yao[1,2,13]

Bipolar disorder is a complex neuropsychiatric disorder that is characterized by intermittent episodes of mania and depression; without treatment, 15% of patients commit suicide[1]. Hence, it has been ranked by the World Health Organization as a top disorder of morbidity and lost productivity[2]. Previous neuropathological studies have revealed a series of alterations in the brains of patients with bipolar disorder or animal models[3], such as reduced glial cell number in the prefrontal cortex of patients[4], upregulated activities of the protein kinase A and C pathways[5–7] and changes in neurotransmission[8–11]. However, the roles and causation of these changes in bipolar disorder have been too complex to exactly determine the pathology of the disease. Furthermore, although some patients show remarkable improvement with lithium treatment for yet unknown reasons, others are refractory to lithium treatment. Therefore, developing an accurate and powerful biological model for bipolar disorder has been a challenge. The introduction of induced pluripotent stem-cell (iPSC) technology has provided a new approach. Here we have developed an iPSC model for human bipolar disorder and investigated the cellular phenotypes of hippocampal dentate gyrus-like neurons derived from iPSCs of patients with bipolar disorder. Guided by RNA sequencing expression profiling, we have detected mitochondrial abnormalities in young neurons from patients with bipolar disorder by using mitochondrial assays; in addition, using both patch-clamp recording and somatic Ca^{2+} imaging, we have observed hyperactive action-potential firing. This hyperexcitability phenotype of young neurons in bipolar disorder was selectively reversed by lithium treatment only in neurons derived from patients who also responded to lithium treatment. Therefore, hyperexcitability is one early endophenotype of bipolar disorder, and our model of iPSCs in this disease might be useful in developing new therapies and drugs aimed at its clinical treatment.

We collected and reprogrammed fibroblasts of six patients with manic type I bipolar disorder (BD) and four unaffected individuals using recombinant Sendai viral vectors expressing the four Yamanaka factors[12] (Extended Data Fig. 1a–c). On the basis of a series of quality control examinations, we selected two clones from each individual for functional experiments (Extended Data Fig. 1d–j). The hippocampus of patients with BD often shows a reduced number of neurons[13,14], indicating that hippocampal neurons probably exhibit cellular phenotypes of BD. We therefore differentiated the iPSCs into hippocampal dentate gyrus (DG) granule cell-like neurons using a newly reported protocol[15] (Fig. 1a, b). More than 80% of the differentiated cells were VGLUT1-positive glutamatergic neurons, most of which were DG granule cell-like neurons that could be identified by a Prox1 promoter-driven lentiviral vector expressing enhanced green fluorescent protein (eGFP) (Prox1::eGFP) or an anti-Prox1 antibody[15]; only 2–7% cells were GABAergic (γ-aminobutyric-acid-releasing) neurons (Fig. 1c, d and Extended Data Fig. 2). Normal and BD neurons showed similar densities of glutamatergic and GABAergic synapses (Fig. 1e, f).

To assess the genetic factors that distinguish patients with BD from healthy people, we performed total RNA sequencing (RNA-seq) analysis to compare the gene expression profiles between 3-week-old BD and normal neurons (Fig. 1g). Compared with normal neurons, 45 genes were significantly differentially expressed in the diseased neurons, with a P value adjusted for false discovery rate (P_{adj}) of ≤ 0.1. Strikingly, we found that the expression of multiple mitochondria genes was significantly enhanced in the BD neurons (Fig. 1h). Clinical studies have revealed that people with mitochondrial cytopathies harbour a high risk of psychiatric disorders, including BD[16,17]. Hence, we investigated the mitochondrial function in young DG-like neurons by measuring the mitochondrial membrane potential (MMP) using the JC-1 assay (Fig. 1i). Flow cytometry analysis revealed that BD neurons showed increased red/green ratios, indicative of enhanced mitochondrial function (Fig. 1j, k and Extended Data Fig. 3a), a finding that is in line with the upregulated mitochondrial gene expression. We next measured the size of neuronal mitochondria, which was represented by the area of DsRed2-mito puncta (Fig. 1l). Compared with normal neurons, the young BD neurons had smaller mitochondria (Fig. 1m, n and Extended Data Fig. 3b). It has been suggested that microtubule-based transport of mitochondria interacts with their dynamics (fusion/fission; morphology or size) and MMP[18]. Moreover, neuronal activity is increased with fast mitochondrial transport and vice versa[19]. Thus, the smaller size and higher MMP of mitochondria in BD neurons probably assist their transport, which might lead to enhanced neuronal activity.

To explore the possible fold change of the mitochondrial alterations in the BD neurons, we expanded our standard of RNA-seq analysis to $|\log_2(\text{fold change})| \geq 1$ and $P \leq 0.05$. We found that 1,005 genes were significantly upregulated and 153 genes were downregulated in the

[1]State Key Laboratory of Membrane Biology, Tsinghua-Peking Joint Center for Life Sciences, McGovern Institute for Brain Research, School of Life Sciences, Tsinghua University, Beijing 100084, China. [2]The Salk Institute for Biological Studies, Laboratory of Genetics, La Jolla, California 92037, USA. [3]The Salk Institute for Biological Studies, Stem Cell Core, La Jolla, California 92037, USA. [4]Key Laboratory of Genomic and Precision Medicine, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, China. [5]Department of Psychiatry, Indiana University, Indianapolis, Indiana 46202, USA. [6]Department of Psychiatry, Case Western Reserve University, Cleveland, Ohio 44106, USA. [7]Department of Psychiatry, University of Bergen, Bergen 5020, Norway. [8]Department of Psychiatry, VA San Diego Healthcare System, La Jolla, California 92151, USA. [9]Department of Psychiatry, University of California San Diego, La Jolla, California, 92093, USA. [10]Department of Psychiatry, Johns Hopkins University, Baltimore, Maryland 21218, USA. [11]Department of Psychiatry, Dalhousie University, Halifax, Nova Scotia, B3H2E2, Canada. [12]Department of Psychiatry, Mount Sinai School of Medicine, New York, New York 10029, USA. [13]Jiangsu Collaborative Innovation Center for Language Ability, Jiangsu Normal University, Xuzhou 221009, China.

*These authors contributed equally to this work.

†Lists of participants and their affiliations appear in the Supplementary Information.

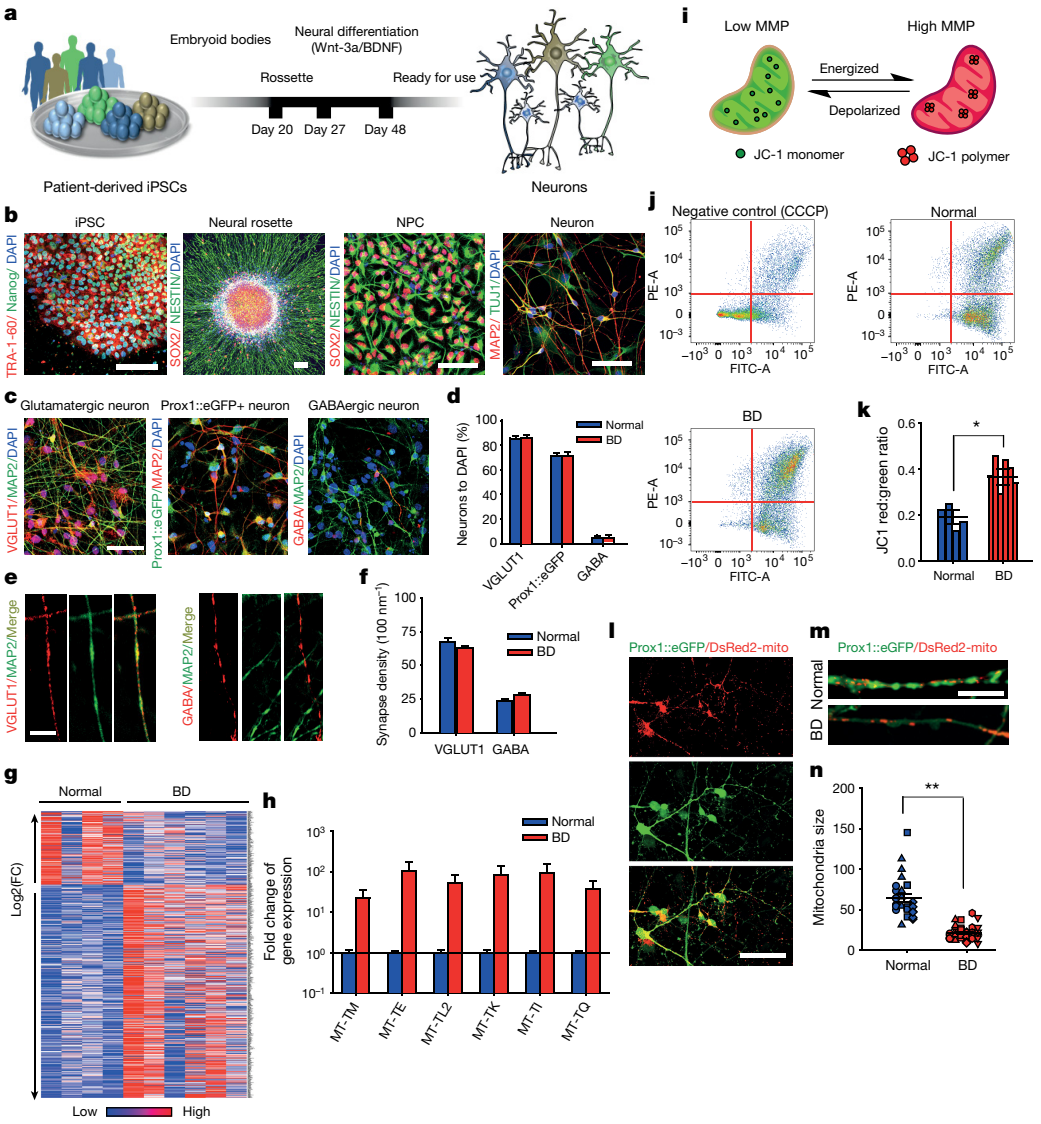


Figure 1 | Hippocampal DG granule cell-like neurons derived from patients with BD show gene expression and mitochondrial abnormalities. **a**, Schematic: generation of DG-like neurons from BD iPSCs. **b**, Immunostainings of iPSCs for TRA-1–60 and Nanog, neural rosettes and neural progenitor cells for SOX2 and Nestin, and neurons for MAP2 and TUJ1. **c**, Immunostainings of neurons labelled with VGLUT1, MAP2, Prox1::eGFP and GABA. Scale bars, 50 µm for **b** and **c**. **d**, Quantification of VGLUT1-positive glutamatergic neurons (normal, $n = 8$; BD, $n = 12$ lines), Prox1::eGFP-positive DG-like neurons (normal, $n = 8$; BD, $n = 12$ lines) and GABAergic neurons (normal, $n = 4$; BD, $n = 12$ lines). **e**, Immunostaining of dendritic glutamatergic synapses and axonal GABAergic synapses. Scale bar, 5 µm. **f**, Quantification of glutamatergic and GABAergic synapse densities (VGLUT1: normal, $n = 30$ neurons from 8 lines; BD, $n = 78$ from 12 lines. GABA: normal, $n = 30$ from 6 lines; BD, $n = 88$ from 12 lines). **g**, Heat map of differential gene expression in normal and BD neurons. **h**, Bar graph summarizing differential expression of mitochondrial genes in BD and normal neurons. **i**, Schematic rationale of JC-1. **j**, JC-1 flow cytometry graphs showing that, as a control, CCCP diminishes neuronal MMP and that BD neurons have elevated MMP. **k**, Quantification of elevated MMP in BD neurons compared with normal (normal, $n = 8$ lines from 4 subjects; BD, $n = 12$ lines from 6 subjects). **l**, **m**, Neurons expressing DsRed2-mito and Prox1::eGFP. Scale bars, 50 µm (**l**) and 20 µm (**m**). **n**, DsRed2-mito puncta sizes reduced in BD neurons. Identical symbols indicate same subject (normal, $n = 29$ cells from 8 lines; BD, $n = 39$ from 12 lines). Student's t-test, $^*P < 0.05$; $^{**}P < 0.001$. Bars, mean ± s.e.m.

BD neurons compared with controls (Fig. 1g). Kyoto Encyclopedia of Genes and Genomes (KEGG) analysis revealed that the Ca^{2+} signalling and neuroactive ligand–receptor interaction pathways were significantly altered (Supplementary Table 1). Gene ontology (GO) analysis suggested that genes involved in the protein kinase A and C (PKA/PKC) signalling pathways and the action potential (AP) firing system were upregulated (Fig. 2a and Supplementary Tables 2 and 3). These observations were verified using quantitative reverse transcription PCR (qRT–PCR) analysis on representative genes (Fig. 2b). Given the facts that enhanced mitochondrial function provides an extra energy resource for AP firing and that upregulation of the PKA/PKC pathways can enhance AP firing[20–22], it is likely that AP firing efficiency is altered in BD.

We therefore performed patch-clamp recording experiments to compare the AP firing patterns of BD and normal iPSC-derived, 3-week-old Prox1::eGFP-labelled DG-like neurons, which had normal synaptic transmission (Fig. 2c, d). Compared with the control neurons, BD neurons exhibited greater activation of Na^+ channels, lower AP threshold and greater values of evoked AP number and maximal AP amplitude (Fig. 2e–k and Extended Data Fig. 3c–f). Further analysis of spontaneous AP firing revealed that the BD neurons showed higher AP frequencies (Fig. 2l–n and Extended Data Fig. 3g). These observations are consistent with the RNA-seq and qRT–PCR results. Although an enhanced expression of K^+ channel subunits was also detected, our patch-clamp recording results did not show any significant changes in the K^+ currents (Extended Data Fig. 4). Given the fact that K^+

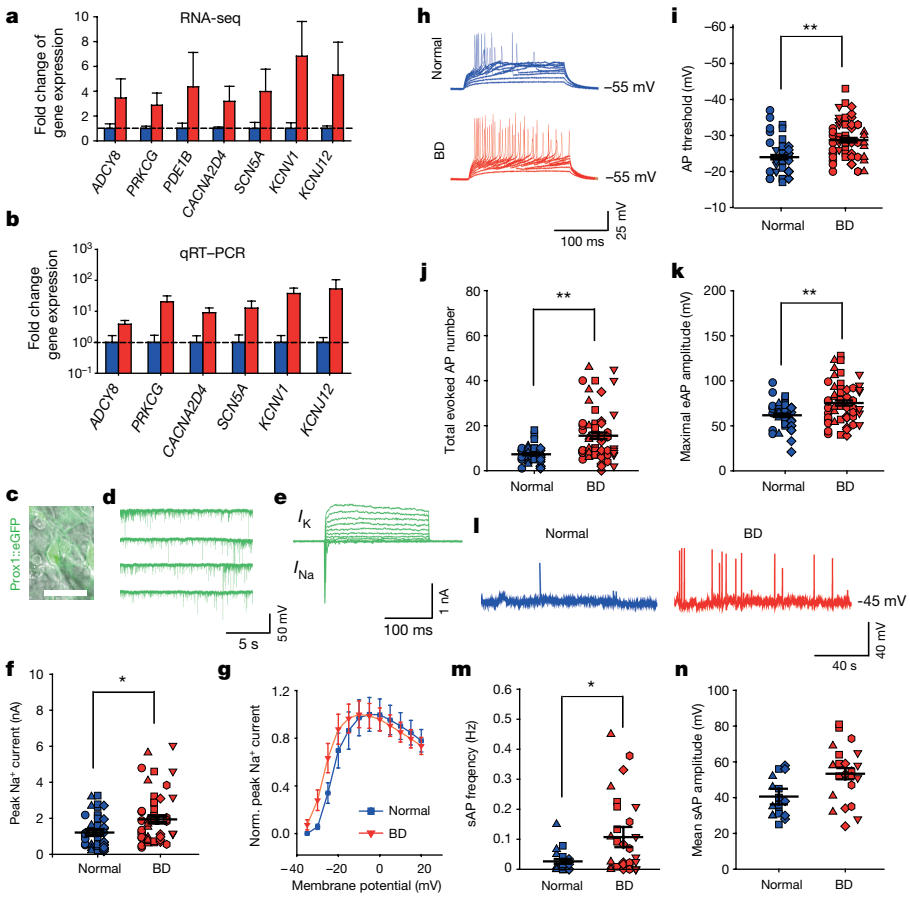


Figure 2 | Hippocampal neurons derived from patients with BD show hyperexcitability. **a**, **b**, Average expression of representative genes involved in the PKA/PKC and AP firing systems revealed by RNA-seq (**a**) and qRT–PCR (**b**) analysis (normal, $n = 4$; BD, $n = 6$ lines). **c**–**e**, Patch-clamp recording on Prox1::eGFP-expressing DG-like neurons (**c**) showed spontaneous postsynaptic currents (**d**) and Na^+/K^+ currents (**e**). Scale bar, 20 µm. **f**, Average peak values of Na^+ currents during stepwise depolarization (normal, $n = 40$ neurons from 8 lines; BD, $n = 52$ from 12 lines). **g**, Normalized average Na^+ currents at different membrane potentials. **h**–**k**, Sample trace (**h**), average firing threshold (**i**), average total number (**j**) and maximal amplitude (**k**) of APs evoked during 300 ms stepwise depolarization. Identical symbols indicate same subject (for AP threshold: normal, $n = 39$ neurons from 8 lines; BD, $n = 55$ from 12 lines; for total AP number: normal, $n = 39$ from 8 lines; BD, $n = 58$ from 12 lines; for maximal amplitude: normal, $n = 39$ from 8 lines; BD, $n = 57$ from 12 lines). **l**–**n**, Sample trace (**l**), average frequency (**m**) and mean amplitude (**n**) of spontaneous APs (for AP frequency: normal, $n = 29$ neurons from 6 lines; BD, $n = 30$ from 8 lines). Student's *t*-test, $*P < 0.05$; $**P < 0.001$. Bars, mean ± s.e.m.

channel subunits are often overexpressed in epilepsy as a compensatory response, the upregulated K^+ channel subunit expression in the BD neurons is probably a homeostatic change by which the neurons attempt to counteract their hyperactivity.

To test the suitability of our BD iPSC model for studying new clinical therapies and drugs, we next set out to investigate the consistency of the hyperactivity phenotype shown in the BD patient-derived neurons with the clinical defects of the patients. Clinically, lithium (Li) has been widely used to treat BD mania. In our study, the recruited subjects included three Li-responsive (LR) and three Li non-responsive (NR) patients (Supplementary Table 4). LR and NR patient-derived neurons exhibited similar percentages of Prox1-positive DG-like cells (Extended Data Fig. 5a, b). Hence, while we were comparing the electrophysiological activity of the BD and normal cells, we also investigated the effects of Li on the activity of the two subgroups of BD neurons in parallel, using 1-week chronic application of 1 mM LiCl. In 3-week-old neurons derived from LR patients, Li significantly reduced Na^+/K^+ currents (Fig. 3a–c), the total number of evoked APs (Fig. 3d, e) and the frequency of spontaneous APs (Fig. 3f, g), whereas the AP amplitudes and threshold remained unaffected (Extended Data Fig. 5c–e). In contrast, Li failed to induce any obvious changes in the NR neurons; however, NR neurons could be affected by the anti-epileptic drug lamotrigine (Extended Data Fig. 6). These results indicated that the hyperactivity of the DG-like neurons that were derived from clinically Li-responsive patients could be selectively diminished by Li treatment. Therefore, the neuronal hyperactivity revealed by our BD iPSC model is directly associated with the clinical symptom of mania in the patients with BD.

To explore the mechanisms that might underlie the Li-caused reduction of neuronal activity in the LR neurons, we performed RNA-seq analysis of Li-treated neurons. We found that, in the NR neurons, 40 genes were changed by the Li treatment; in sharp contrast, 560 genes in the LR neurons were significantly affected, of which 238 were upregulated and 322 were downregulated (Fig. 3h, i). Hence, Li can specifically affect the gene expression profiles of the LR neurons. Further analysis revealed that Li rescued 84 genes in the LR neurons to varying degrees, including the gene(s) that are probably key for the BD pathology and Li response, and thus could potentially be used to develop DNA predictor systems. Of these 84 genes, those involved in the PKA/PKC pathways and AP firing, such as *PDE11A*, *PRKCH*, *PTPRB* and *SCN11A*, as well as multiple mitochondria-related genes, were significantly downregulated, and the Na^+/K^+ ATPase pathway gene *NKAIN* was upregulated (Fig. 3j, k), indicating the attenuation of the PKA/PKC pathways, AP firing system, and mitochondrial functions. Indeed, Li partly rescued mitochondrial dysfunction by increasing the mitochondria size in 3-week-old LR neurons, whereas the MMP remained unaffected (Fig. 3l–n). It thus appears conceivable that Li diminishes hyperactivity of the LR neurons through reversing aberrant gene expression related to these pathway(s). In addition, we found that the expression of K^+ subunits (*KCNA1*, *KCNJ12*) was also significantly downregulated in response to the Li treatment (Fig. 3j), probably because of a neuronal homeostatic response to the loss of neuronal activity.

We next tested whether the enhanced excitability of single neurons could generate neural network hyperactivity through assaying somatic Ca^{2+} transient events with a calcium indicator, Fluo 4-AM. Ca^{2+} events were abolished by tetrodotoxin (TTX) (Fig. 4a, b), indicating that they represent APs spreading over the neural network[23–25]. As Prox1-expressing DG-like neurons accounted for approximately 80% of all neurons in the culture dish (Fig. 1c, d), Ca^{2+} imaging of synapsin promoter-driven lentiviral vector expressing DsRed (Syn::DsRed)-labelled neurons was able to monitor the activity of the granule cells. Compared with the normal group, the BD LR and NR neural networks both showed a significantly higher frequency of Ca^{2+} events (Fig. 4c–e). In the LR neural network, Li application resulted in a remarkable reduction both in the Ca^{2+} event frequency and in the

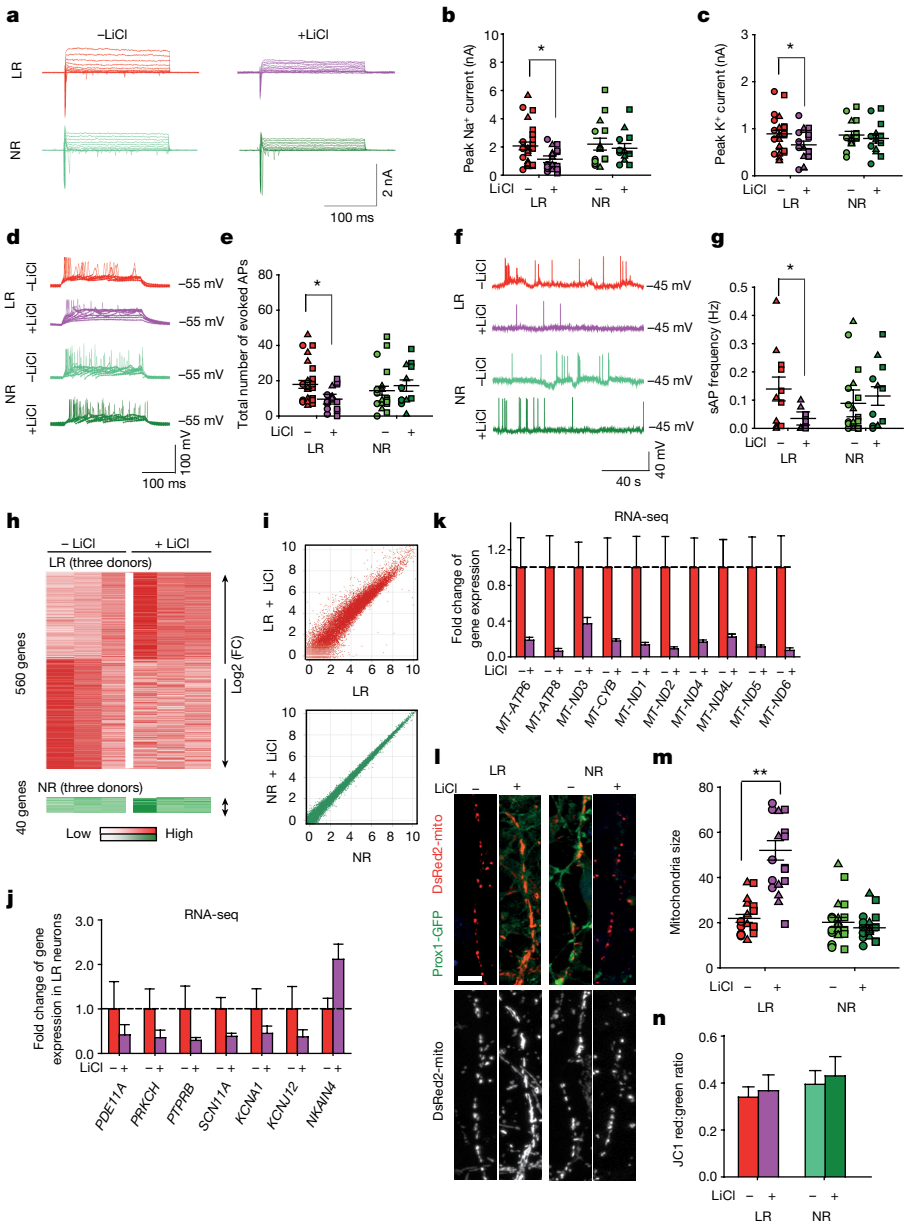


Figure 3 | Li rescues hyperexcitability in hippocampal neurons derived from iPSCs of patients with BD. **a**, Na^+/K^+ currents recorded from BD LR and NR neurons with and without Li. **b**, **c**, Effects of Li on average peaks of Na^+ currents (**b**) and K^+ currents (**c**) in the LR and NR neurons. Identical symbols indicate same subject (LR without Li, $n = 26$ neurons from 5 lines; with Li, $n = 19$ from 5 lines). **d**, Representative traces of APs evoked during 300 ms stepwise depolarization periods. **e**, Scatter graph showing Li-induced decrease in the average total AP number of the LR neurons (LR without Li treatment, $n = 27$ neurons from 5 lines; with Li treatment, $n = 18$ from 5 lines). **f**, Representative traces of spontaneous APs. **g**, Spontaneous AP firing frequency in Li-treated LR neurons (LR without Li, $n = 11$ neurons from 3 lines; with Li, $n = 10$ from 3 lines). **h**, **i**, Heat maps (**h**) and MA plots (**i**) showing effects of Li treatment on gene expression in LR and NR neurons. **j**, **k**, Effects of Li on the average expression of representative PKA/PKC/AP (**j**) and mitochondrial genes (**k**) in the LR neurons (with Li, $n = 3$; without Li, $n = 3$ lines). **l**, **m**, Sample images of neurons (**l**) and bar graph (**m**) showing the effects of Li treatment on mitochondria morphology. Scale bar, 10 μm ($n = 19$ neurons from 6 lines). **n**, No effects of Li treatment on MMP of the BD neurons ($n = 6$ lines for each group). Student's *t*-test, $*P < 0.05$; $**P < 0.001$. Bars, mean ± s.e.m.

percentage of signalling neurons (Fig. 4d, e and Extended Data Fig. 3g), whereas the BD NR network was unaffected. Interestingly, normal neurons did not show any obvious changes either (Extended Data Fig. 7). In addition, we observed that this hyperexcitability would reverse when the diseased neurons became old (Extended Data Fig. 8), which may represent an early sign of the reported loss of mature hippocampal neurons in the BD brain and/or might be associated with the transition of the patients from mania into depression[13,14].

Previously, hyperexcitability had been observed in the ventral tegmental area dopaminergic neurons and hippocampal DG neurons of BD animal models, and thus was thought to be an endophenotype of this disease[26,27]. However, it remained unclear whether this phenotype could represent the clinical symptoms of BD. In the present study, we generated iPSCs from the fibroblasts of clinically diagnosed patients with BD and demonstrated that 3-week-old diseased neurons derived from iPSCs exhibited significantly upregulated neuronal activity. Importantly, we found that treating neurons with Li selectively diminished this abnormality only in neurons derived from those patients who were responsive to clinical Li administration. Notably, in a neuronal model of schizophrenia generated by the identical approach[15], we did not observe hyperactivity in the diseased neurons (data not shown). Therefore, our findings indicated that neuronal hyperexcitability is specifically associated with the clinical symptoms of patients with BD.

As indicated earlier, patients with BD differ in response to Li; a subset of patients has a very robust response with excellent control of symptoms whereas others do not. This variability in response leads frequently to many years of trial-and-error efforts to identify the optimal medication. Recognition of this differential responsiveness to Li may lead not only to novel treatments but also to DNA or other biomarker predictors of response that can accelerate treatment optimization and provide precision medicine in psychiatry. Using neuronal hyperactivity and Li responsiveness as two indices, we detected correlated changes in the PKA/PKC/AP and mitochondria genes in the BD neurons, indicating that these pathways might be related to neuronal hyperexcitability. Further investigations will be necessary to determine whether mitochondrial alterations and/or PKA/PKC/AP gene expression changes represent a cause or a consequence of the observed hyperexcitability phenotype and to whether the reversal of hyperexcitability represents further progression of the disease.

In summary, the cell-autonomous findings revealed by our BD neuronal model based on iPSC technology represent an important first step in understanding the pathophysiology of BD, improving

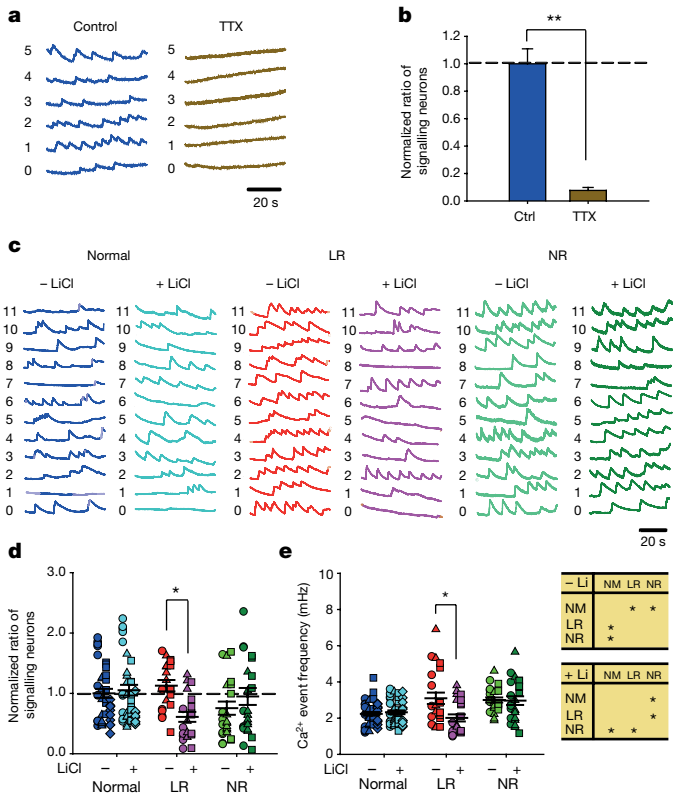


Figure 4 | Somatic Ca^{2+} imaging analysis reveals hyperactivity in the neural network formed by the BD iPSC-derived neurons. a, **b**, Sample traces (**a**) and bar graph (**b**) showing neuronal Ca^{2+} transients abolished by tetrodotoxin ($n = 10$ images). **c**, Representative Ca^{2+} traces in normal, BD LR and NR neurons. **d**, Effects of Li treatment on the average ratio of neurons exhibiting Ca^{2+} events. Identical symbols indicate the same subject ($n = 23$ images from 6 lines). **e**, Scatter graph (left) and analysis of variance (ANOVA) (right) showing the average Ca^{2+} event frequencies in normal and BD neurons treated with Li (normal, $n = 43$ images from 8 lines; LR, $n = 23$ from 6 lines; NR, $n = 23$ from 6 lines). Student's t-test (**b**) and ANOVA (**d**, **e**), $^*P < 0.05$; $^{**}P < 0.001$. Bars, mean ± s.e.m.

diagnosis and perhaps developing novel therapeutics for treatment of the disease.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 1 July 2014; accepted 26 August 2015.
Published online 28 October 2015.

1. Sharma, R. & Markar, H. R. Mortality in affective disorder. *J. Affect. Disord.* **31,** 91–96 (1994).
2. Dusetzina, S. B. *et al.* Treatment use and costs among privately insured youths with diagnoses of bipolar disorder. *Psychiatr. Serv.* **63,** 1019–1025 (2012).
3. Martinowich, K., Schloesser, R. J. & Manji, H. K. Bipolar disorder: from genes to behavior pathways. *J. Clin. Invest.* **119,** 726–736 (2009).
4. Andreazza, A. C. & Young, L. T. The neurobiology of bipolar disorder: identifying targets for specific agents and synergies for combination treatment. *Int. J. Neuropsychopharmacol.* **17,** 1039–1052 (2014).
5. Chang, A., Li, P. P. & Warsh, J. J. cAMP-Dependent protein kinase (PKA) subunit mRNA levels in postmortem brain from patients with bipolar affective disorder (BD). *Brain Res. Mol. Brain Res.* **116,** 27–37 (2003).
6. Bezchlibnyk, Y. & Young, L. T. The neurobiology of bipolar disorder: focus on signal transduction pathways and the regulation of gene expression. *Can. J. Psychiatry* **47,** 135–148 (2002).
7. Wang, H. & Friedman, E. Increased association of brain protein kinase C with the receptor for activated C kinase-1 (RACK1) in bipolar affective disorder. *Biol. Psychiatry* **50,** 364–370 (2001).
8. Berk, M. *et al.* Dopamine dysregulation syndrome: implications for a dopamine hypothesis of bipolar disorder. *Acta Psychiatr. Scand. Suppl.* **434,** 41–49 (2007).
9. Mahmood, T. & Silverstone, T. Serotonin and bipolar disorder. *J. Affect. Disord.* **66,** 1–11 (2001).
10. Scarr, E., Pavey, G., Sundram, S., MacKinnon, A. & Dean, B. Decreased hippocampal NMDA, but not kainate or AMPA receptors in bipolar disorder. *Bipolar Disord.* **5,** 257–264 (2003).
11. Du, J., Quiroz, J., Yuan, P., Zarate, C. & Manji, H. K. Bipolar disorder: involvement of signaling cascades and AMPA receptor trafficking at synapses. *Neuron Glia Biol.* **1,** 231–243 (2004).
12. Takahashi, K. *et al.* Induction of pluripotent stem cells from adult human fibroblasts by defined factors. *Cell* **131,** 861–872 (2007).
13. Bertolino, A. *et al.* Neuronal pathology in the hippocampal area of patients with bipolar disorder: a study with proton magnetic resonance spectroscopic imaging. *Biol. Psychiatry* **53,** 906–913 (2003).
14. Deicken, R. F., Pegues, M. P., Anzalone, S., Feiwell, R. & Soher, B. Lower concentration of hippocampal *N*-acetylaspartate in familial bipolar I disorder. *Am. J. Psychiatry* **160,** 873–882 (2003).
15. Yu, D. X. *et al.* Modeling hippocampal neurogenesis using human pluripotent stem cells. *Stem Cell Rep.* **2,** 295–310 (2014).
16. Fattal, O., Link, J., Quinn, K., Cohen, B. H. & Franco, K. Psychiatric comorbidity in 36 adults with mitochondrial cytopathies. *CNS Spectr.* **12,** 429–438 (2007).
17. Marazziti, D. *et al.* Psychiatric disorders and mitochondrial dysfunctions. *Eur. Rev. Med. Pharmacol. Sci.* **16,** 270–275 (2012).
18. Chen, H. & Chan, D. C. Mitochondrial dynamics–fusion, fission, movement, and mitophagy–in neurodegenerative diseases. *Hum. Mol. Genet.* **18** (Suppl. R2), R169–R176 (2009).
19. Sun, T., Qiao, H., Pan, P. Y., Chen, Y. & Sheng, Z. H. Motile axonal mitochondria contribute to the variability of presynaptic strength. *Cell Reports* **4,** 413–419 (2013).
20. Kobayashi, M., Sasabe, T., Shiohama, Y. & Koshikawa, N. Activation of α_1-adrenoceptors increases firing frequency through protein kinase C in pyramidal neurons of rat visual cortex. *Neurosci. Lett.* **430,** 175–180 (2008).
21. Szulczyk, B., Książek, A., Ładno, W. & Szulczyk, P. Effect of dopamine receptor stimulation on voltage-dependent fast-inactivating Na^+ currents in medial prefrontal cortex (mPFC) pyramidal neurons in adult rats. *Acta Neurobiol. Exp. (Warsz.)* **72,** 351–364 (2012).
22. Yuan, L. L., Adams, J. P., Swank, M., Sweatt, J. D. & Johnston, D. Protein kinase modulation of dendritic K^+ channels in hippocampus involves a mitogen-activated protein kinase pathway. *J. Neurosci.* **22,** 4860–4868 (2002).
23. Marchetto, M. C. *et al.* A model for neural development and treatment of Rett syndrome using human induced pluripotent stem cells. *Cell* **143,** 527–539 (2010).
24. Yuste, R., MacLean, J., Vogelstein, J. & Paninski, L. Imaging action potentials with calcium indicators. *Cold Spring Harb. Protoc.* **2011,** 985–989 (2011).
25. Grewe, B. F., Langer, D., Kasper, H., Kampa, B. M. & Helmchen, F. High-speed *in vivo* calcium imaging reveals neuronal network activity with near-millisecond precision. *Nature Methods* **7,** 399–405 (2010).
26. Coque, L. *et al.* Specific role of VTA dopamine neuronal firing rates and morphology in the reversal of anxiety-related, but not depression-related behavior in the ClockΔ19 mouse model of mania. *Neuropsychopharmacology* **36,** 1478–1488 (2011).
27. Hagihara, H., Takao, K., Walton, N. M., Matsumoto, M. & Miyakawa, T. Immature dentate gyrus: an endophenotype of neuropsychiatric disorders. *Neural Plast.* **2013,** 318596 (2013).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank the patients who participated in this study. We thank M. Ku for help in the RNA-seq analysis, and L. McHenry, D. Lisuk and C. Bardy for help with the somatic Ca^{2+} imaging experiments. We thank L. Moore, E. Mejia, B. Miller, R. Wright, T. Berggren and S. Lu for technical assistance. We also thank C. O'Connor for help on flow cytometry. This work was supported by the National Natural Science Foundation of China (grant numbers 31471020, 31161120358, 31123004), the National Basic Research Program of China (2015CB910603, 2011CB510106), the Open Project of Key Laboratory of Genomic and Precision Medicine, Chinese Academy of Sciences, by the Engmann Foundation, the JPB Foundation, the Helmsley Trust, the Mather's Foundation, the Glenn Foundation for Aging Research, by National Institute of Health grant MH106056 (K.J.B.), New York Stem Cell Foundation – Robertson Award (K.J.B.), and by grants/contracts to J.R.K. from the National Institute of Mental Health (U01 MH92758) supporting the Pharmacogenomics of Bipolar Disorder Study and from the Department of Veterans Affairs (5I01CX000363). K.J.B. is a New York Stem Cell Foundation - Robertson Investigator. J.Y. is an Investigator of the Young Thousand Talents Program of China.

Author Contributions J.M., Q.W.W., K.E.D., L.B., K.J.B., T.J.E. and J.Y. conducted the iPSC reprogramming and differentiation experiments. Q.W.W., B.Y. and J.Y. conducted the electrophysiological recording experiments. J.M., Q.W.W., Y.Z., S.S. and J.Y. conducted immunostaining experiments. D.X.Y., J.M., Q.W.W., B.Y. S.T.S. and J.Y. conducted Ca^{2+} imaging experiments. Y.K., Q.W.W. and J.Y. performed mitochondrial assays. J.M., S.P., B.Y., J.Z., Y.Z., S.M. and J.Y. conducted RNA-seq and qRT-PCR analysis. J.R.K., J.I.N., J.R.C., K.J.O., M.J.M., P.P.Z., M.A., C.M.N. and the Pharmacogenomics of Bipolar Disorder Study designed and conducted the clinical trial and provided samples from patients. J.Y. designed the experiments with F.H.G. and wrote the manuscript with J.M. and Q.W.W.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to J.Y. (jyao@mail.tsinghua.edu.cn), F.H.G. (gage@salk.edu) or J.R.K. (jkelsoe@ucsd.edu).

METHODS

Subjects. Patients with type I BD were participants in one of two prospective clinical trials of Li monotherapy for identifying genetic variants predictive of good Li response. These studies included one of Li response in veterans conducted at the University of California, San Diego; the other was the Pharmacogenomics of Bipolar Disorder Study (clinical trial number NCT01272531; Supplementary Table 5). All subjects provided written informed consent and all procedures were approved by local human subjects committees. Subjects were initially screened for eligibility and diagnoses determined using the Diagnostic Interview for Genetic Studies. Subjects were started on Li. Over 4 months they were titrated up to a therapeutic level (1.0 meq dl^{-1}) as tolerated, as other psychotropic medications were gradually discontinued. Subjects were seen at 2- to 4-week intervals and rated for mood symptoms. After 4 months, subjects with a Clinical Global Impressions Scale score of 3 or less, reflecting only mild symptoms, were declared responders, whereas other subjects were deemed non-responders. Li responders were then followed on Li monotherapy for up to 2 years. The responders remained stable for an average of 23 months on Li monotherapy, whereas the non-responders failed to remit from their index episode after an average trial of Li of 3 months. All subjects were white males. The characteristics of these subjects are detailed in Supplementary Table 4. Four-millimetre punch skin biopsies were obtained under sterile conditions from a few centimetres below the iliac crest.

iPSC reprogramming and neuron differentiation. BD and normal iPSCs were derived from fibroblasts using a Cyto-Tune Sendai reprogramming kit (Invitrogen) according to the manufacturer's instructions. All iPSCs were characterized as previously described[28]. iPSC colonies were cultured on Matrigel-coated dishes (BD Biosciences) using mTeSR1 medium (StemCell Technologies). Embryoid bodies were formed by mechanical dissociation of iPSC colonies using collagenase and plating onto low-adherence dishes in DMEM/F12 (Invitrogen) supplemented with N2 and B27. For embryoid body differentiation, floating embryoid bodies were treated with DKK1 (0.5 μg ml^{-1}), SB431542 (10 μm), noggin (0.5 μg ml^{-1}) and cyclopamine (1 μm) for 20 days. To obtain neural progenitor cells, embryoid bodies were plated onto polyornithine/laminin (Sigma)-coated dishes in DMEM/F12 plus N2 and B27. Rosettes were manually collected and dissociated with accutase (Chemicon) after 1 week and plated onto laminin-coated dishes in neural progenitor cell media (DMEM/F12, 1× N2, 1× B27 (Invitrogen), 1 μg ml^{-1} laminin and 20 ng ml^{-1} FGF2 (Invitrogen)). To obtain mature neurons, neural progenitor cells were differentiated in DMEM/F12 supplemented with 1× N2, 1× B27, 20 ng ml^{-1} BDNF (Peprotech), 1 mM dibutyrl-cyclicAMP (Sigma), 200 nM ascorbic acid (Sigma), 1 μg ml^{-1} laminin and 620 ng ml^{-1} Wnt3a (R&D) for 3 weeks. Wnt3a was removed after 3 weeks. All cells used in the present study were verified as free from mycoplasma contamination.

Generation of lentivirus. Lentivirus was packaged in 293T HEK cells grown in DMEM/F12 (Invitrogen) supplemented with 10% FBS (Gemini). The 293T cells cultured in the 15-cm dish were transfected with a solution consisting of 12.2 μg lentiviral DNA, 8.1 μg MDL-gagpol, 3.1 μg Rev-RSV, 4.1 μg CMV-VSVG, 500 μl of Opti-MEM (Invitrogen) and 110 μl PEI (1 μg ml^{-1}). Medium was changed after 12 h and the virus was harvested at 72 h after transfection.

Immunocytochemistry. Cells were fixed in 4% paraformaldehyde and then permeabilized with 0.25% Triton-X100 in PBS. Cells were then blocked in Tris-Cl buffer solution (TBS) containing 0.25% Triton-X100 and 10% donkey serum for 1 h, followed by incubation with primary antibody overnight at 4 °C. After three washes with TBS, cells were incubated with secondary antibodies for 1 h at room temperature. After three washes with TBS, cells were incubated with DAPI (0.1 μg ml^{-1}, Sigma) for 15 min, followed by three washes with TBS to remove DAPI. Fluorescent signals were detected using a Zeiss 710 laser scanning microscope and images were processed with ZEN 2011, Adobe Photoshop CS5 and ImageJ 1.42 software. The primary antibodies used were mouse anti-TRA-1-60 monoclonal antibody (1:200, Chemicon catalogue number MAB4360), goat anti-Nanog polyclonal antibody (1:200, R&D catalogue number AF1997), goat anti-SOX2 polyclonal antibody (1:250, Santa Cruz catalogue number sc-17320), mouse anti-Nestin monoclonal antibody (1:200, Chemicon catalogue number MAB5326), rabbit anti-TUJ1 polyclonal antibody (1:500, Covance catalogue number PRB-435P), chicken anti-MAP2 polyclonal antibody (1:1,000, Abcam catalogue number ab5392), rabbit anti-VGLUT1 polyclonal antibody (1:200, Synaptic Systems catalogue number ab5392), rabbit anti-GFP antibody (1:500, Invitrogen catalogue number A6455) and rabbit anti-GABA polyclonal antibody (1: 1,000, Sigma catalogue number A2052). The secondary antibodies (Jackson ImmunoResearch Laboratories) used were goat anti-chicken Alexa Fluor 647 (1:500, catalogue number 703-605-155), goat anti-rabbit CY3 (1:500, catalogue number 111-165-003), donkey anti-chicken Dylight 488 (1:500, catalogue number 703-485-155), donkey anti-rabbit CY3 (1:500, catalogue number 711-165-152), donkey anti-rabbit Alexa 488 (1:500, catalogue number 711-545-152), donkey anti-chicken Dylight 549 (1:500, catalogue number 703-505-155), donkey anti-goat Alexa 488 (1:500, catalogue number 705-545-147), donkey anti-mouse CY3 (1:500, catalogue number 715-165-151), donkey anti-goat CY3 (1:500, catalogue number 705-165-147) and donkey anti-mouse Alexa 488 (1:500, catalogue number 715-545-151). All relevant information about the antibodies used in this study, including citation, clone number and antibody validation profile, can be found at the manufacturers' websites.

RNA extraction, PCR and quantitative RT–PCR. Total cellular RNA was extracted from approximately 5×10^6 cells using the RNA-BEE (Qiagen) according to the manufacturer's instructions, and reverse transcription was performed using a High Capacity cDNA Synthesis kit (AB Biosystems). PCR was performed using a GoTAQ PCR kit (Fisher Scientific), and PCR products were analysed using agarose gel electrophoresis. Quantitative PCR was done using SyberGreen (Invitrogen), and the results were analysed using SDS3.2 software for a 7900HT real-time PCR system. The primer sequences used are described in Supplementary Table 6.

Somatic calcium imaging. Three-week-old neurons derived from BD and normal iPSCs were previously infected with a synapsin promoter-driven lentiviral vector expressing DsRed (Syn::DsRed). Cell cultures were washed twice with sterile Krebs HEPES Buffer and incubated with 3 μm Fluo 4-AM (Molecular Probes) in Krebs HEPES Buffer for 40 min at room temperature. Excess dye was removed by washing twice with Krebs HEPES Buffer, and cells were incubated for an additional 20 min to equilibrate the intracellular dye concentration and allow de-esterification. Time-lapse image sequences (×100 magnification) of 3,000 frames were acquired at 28 Hz with a region of 336 pixels × 256 pixels using a Hamamatsu ORCA-ER digital camera (Hamamatsu Photonics) with a 488 nm (FITC (fluorescein isothiocyanate)) filter on an Olympus IX81 inverted fluorescence confocal microscope (Olympus Optical). To assess changes in calcium signalling in response to perturbation of neuronal activity, tetrodotoxin (1 μm) was applied by bath application. Images were acquired with MetaMorph 7.7 software (MDS Analytical Technologies). Images were subsequently processed using ImageJ software and custom written routines in Matlab 7.2 software (Mathworks).

Electrophysiology. Neurons were previously infected with the Prox1::eGFP lentiviral vector. Whole-cell patch-clamp recordings were performed from Prox1::eGFP highlighted DG-like neurons after 3 weeks of differentiation. The bath was constantly perfused with an extracellular solution (128 mM NaCl, 5 mM KCl, 2 mM CaCl2, 30 mM glucose, 1 mM MgCl2 and 25 mM HEPES (pH 7.3)). The recording micropipettes (tip resistance 3–6 MΩ) were filled with internal solution (130 mM K-gluconate, 1 mM EGTA, 2 mM Mg-ATP, 0.3 mM Na-GTP, 5 mM Na-phosphocreatine and 10 mM HEPES (pH 7.3)). Recordings were made using Axopatch 200B or 700B amplifier (Axon Instruments). Signals were filtered at 2 kHz and sampled at 5 kHz. The series resistance was typically <15 MΩ. For voltage-clamp recordings, the membrane potential was held at −70 mV. To record the sodium and potassium currents, cells were depolarized in 5 mV increments. For current-clamp recordings, a hyperpolarized current was injected into the neuron to a membrane potential of −55 mV or −45 mV, depending on the experiments. Step-depolarized currents with identical parameters were injected into normal and BD neurons to elicit APs. All recordings were performed at room temperature and chemicals were purchased from Sigma.

Mitochondrial assay and flow cytometry. To measure mitochondrial size, Prox1::eGFP and DsRed2-mito were co-expressed in DG-like neurons via lentiviral infection. Neurons were fixed in 4% paraformaldehyde and then permeabilized with 0.1% Triton-X100 in TBS. Cells were then blocked in TBS containing 3% donkey serum for 1 h, followed by incubation with DAPI for 15 min. Fluorescence images were acquired using a high-resolution LSM 710 confocal microscope (Carl Zeiss) and were processed with ZEN 2011 software (Carl Zeiss) and Adobe Photoshop CS5 software (Adobe). The size of the mitochondria (DsRed2-mito puncta) was analysed using the Particle Analysis tool in ImageJ software (National Institutes of Health).

For MMP, neurons were incubated with JC-1 dye (Molecular Probes) at 37 °C for 15–30 min[29]. The cells were dissociated into single cells using TrypLE (Invitrogen), washed three times and then resuspended in 1 ml warm PBS. Green and red fluorescence of JC-1 dye was quantitated using BD FACSCanto II flow cytometer (Becton, Dickinson). Histogram plots of green and red fluorescence were created to determine the red/green intensity ratio using FlowJo 10 software (TreeStar).

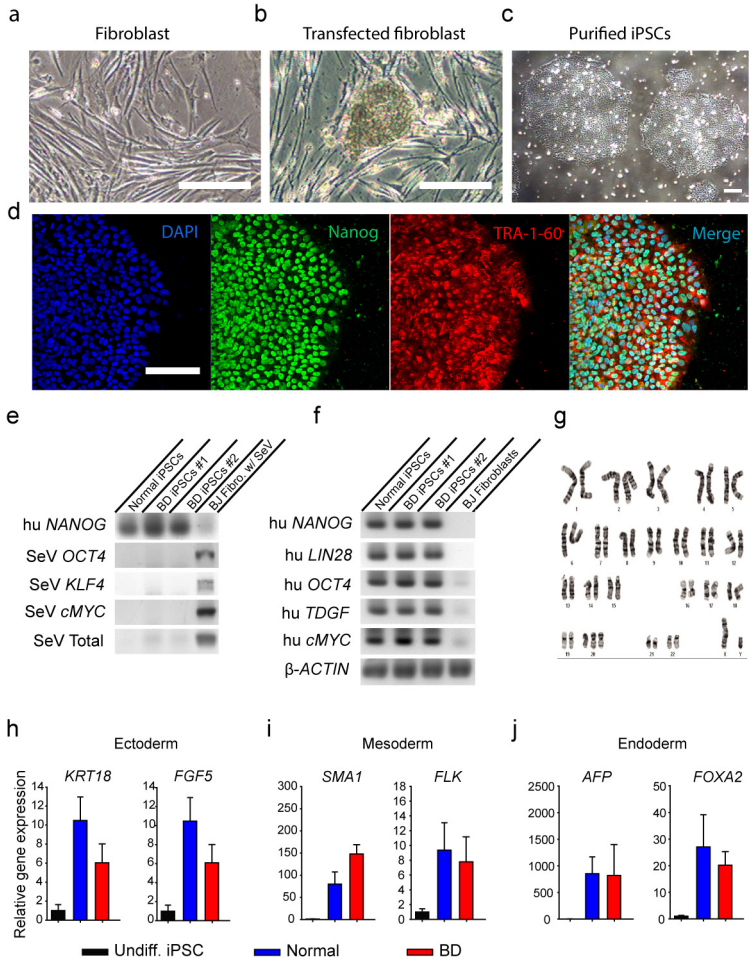
RNA-seq analysis. RNA was prepared into RNA-Seq libraries using an Illumina TruSeq Stranded Total RNA Sample Prep Kit with Ribo-zero Gold (Human/Mouse/Rat) (Illumina). Cytoplasmic and mitochondrial ribosomal RNA was depleted using a Ribo-zero Gold component. Depleted RNA was reverse transcribed into cDNA using SuperScript II reverse transcriptase (Invitrogen). Stranded cDNA sequencing libraries were generated according to Illumina's procedures. Total RNA-Seq libraries were sequenced paired-end 2 × 100 base pairs (bp) using the Illumina HiSeq 2500 platform according to the manufacturer's specifications. Low-quality ends and read-through adaptor sequences were trimmed using Cutadapt, version 1.3. The trimmed reads were mapped to the human genome (hg19/GRCh37) using

STAR, version 2.3.0. The assignment of reads to gene regions was performed by htseq-count, version 0.5.4p5. These raw counts are taken as input for edgeR package for differential gene expression analysis using the exact and paired Student's *t*-test described in the edgeR manual. DAVID (http://david.abcc.ncifcrf.gov/) was used to perform the gene functional annotation analysis. The categories of GO and KEGG pathways were chosen as background databases. All genes of *Homo sapiens* were used as background gene list. The RNA-seq data have been deposited in NCBI's Gene Expression Omnibus under accession number GSE58933. The Prox1::eGFP-positive DG-like neurons showed gene expression similar to the whole differentiation culture (Extended Data Fig. 9).

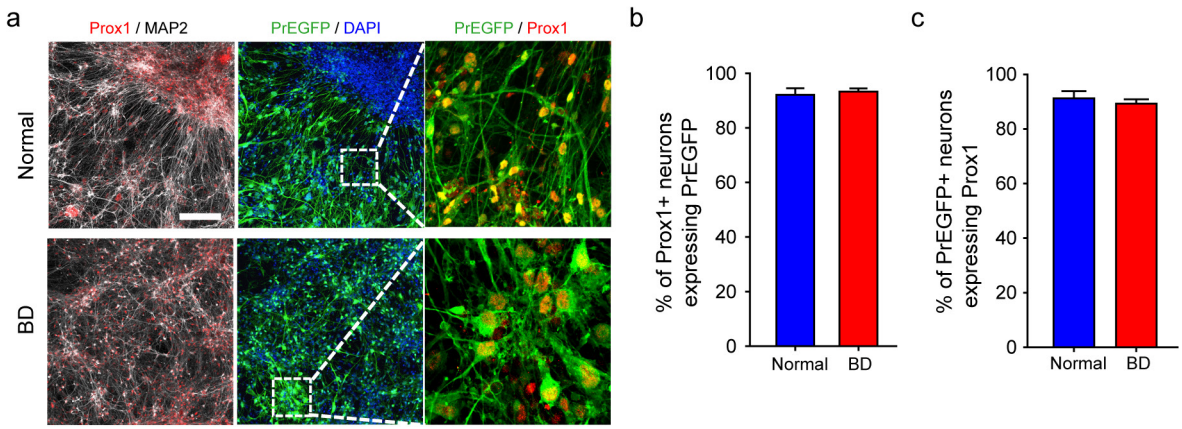
Statistical analysis. No statistical methods were used to predetermine sample size. The experiments were not randomized.

For comparisons of Ca^{2+} imaging results among the normal, LR and NR groups, the difference was assessed using one-way ANOVA followed by Duncan's test; the *P* value was adjusted by Benjamini and Hochberg correction, and an adjusted *P* value <0.05 was considered as significant. For RNA-seq, the data were analysed using the edgeR package[30]. For pairwise comparisons, we used quantile-adjusted conditional maximum likelihood methods. The common dispersion was calculated by using the estimateCommonDisp. The exact test is based on quantile-adjusted conditional maximum likelihood methods. Knowing the conditional distribution of the sum of counts in a group, edgeR computes exact *P* values by summing over all sums of counts that have a probability less than the probability under the null hypothesis of the observed sum of counts. Benjamini and Hochberg's algorithm is used to control the false discovery rate. We performed paired comparisons to detect gene expression changes in response to Li treatment. This is an additive model with the patient as the blocking factor. For all other experiments, a two-tailed unpaired Student's *t*-test was used to determine the statistical significance of observed differences between various conditions. The analysis approaches have been justified as appropriate by previous biological studies, and all data met the criteria of normal distribution. In all experiments, two lines from one patient were prepared, and one or two lines were eventually used for the experiment depending on the status of the cells, such as differentiation and cell density. For most experiments in this study, neurons of all patients and at similar densities were investigated (Extended Data Fig. 10), except that, for recordings of spontaneous AP firing, two patients with BD LR were investigated. The statistical data for each subject are listed in Supplementary Table 7. In the experiments, every cell line had a unique code that could not tell the identity of the subject but could tell which two lines belonged to the same subject, so that the person performing the experiments could use at least one line for each subject without knowing the group category. The collected data were used for statistical analysis without exclusion. All experiments were performed in technical and biological triplicate and were repeated at least three times. The variation within each group of neurons was not pre-estimated and the variation between groups might not be similar. For electrophysiological recording experiments, at least five to six neurons per subject were recorded and statistically analysed without exclusion. For Ca^{2+} imaging and immunostaining experiments, typically four or five view fields per subject containing several hundred neurons were used without exclusion for analysis. For RNA-seq, qRT–PCR and flow cytometry experiments, all cells in one culture were collected for analysis.

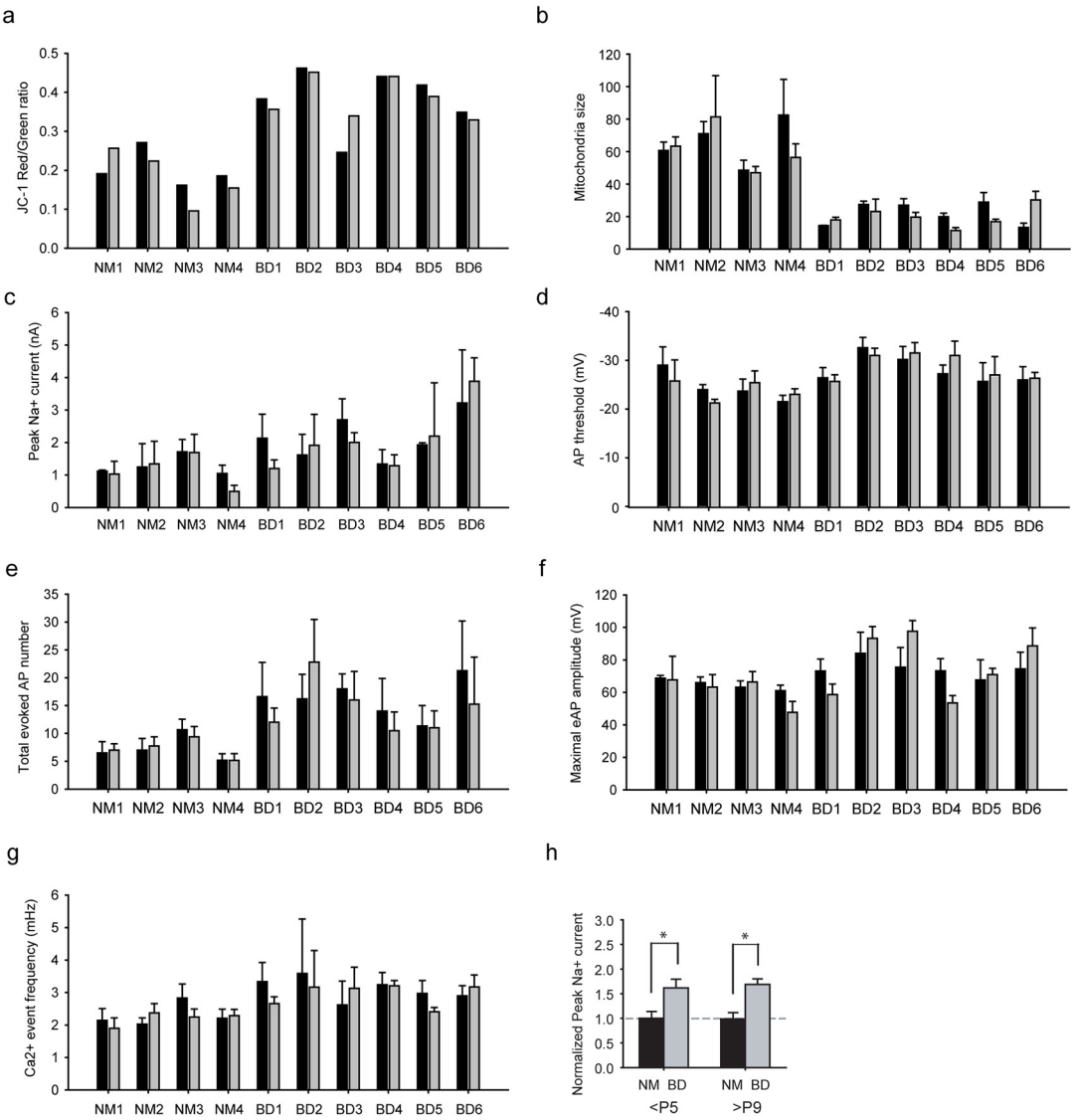
28. Brennand, K. J. *et al.* Modelling schizophrenia using human induced pluripotent stem cells. *Nature* **473,** 221–225 (2011).
29. Brennand, K. *et al.* Phenotypic differences in hiPSC NPCs derived from patients with schizophrenia. *Mol. Psychiatry* **20,** 361–368 (2015).
30. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26,** 139–140 (2010).



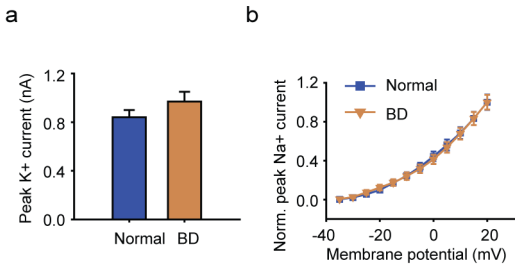
Extended Data Figure 1 | Generation of iPSCs from patients with BD and healthy people. a, Human fibroblasts generated from punch biopsy. **b**, The iPSC colonies appeared after fibroblasts were reprogrammed using the Sendai virus. **c**, Purified iPSC colonies were cultured in Matrigel-coated plate. **d**, Immunostaining of iPSCs with DAPI and pluripotency markers Nanog and TRA-1-60. **e**, RT–PCR results showing that the introduced Sendai virus genes were cleared from the generated iPSCs. **f**, RT–PCR results showing that the generated iPSCs expressed human pluripotency markers *NANOG*, *LIN28*, *OCT4*, *TDGF* and *cMYC*. **g**, Representative karyotyping image of generated iPSCs showing normal chromosomal structure. **h**–**j**, Bar graphs of quantitative RT–PCR showing that the iPSCs can randomly differentiate into cells expressing the markers for endoderm, mesoderm and ectoderm. Data are representative for a total of 20 iPS cell lines from 10 patients (2 clones per patient). Scale bar, 50 μm. Bars, mean ± s.e.m.



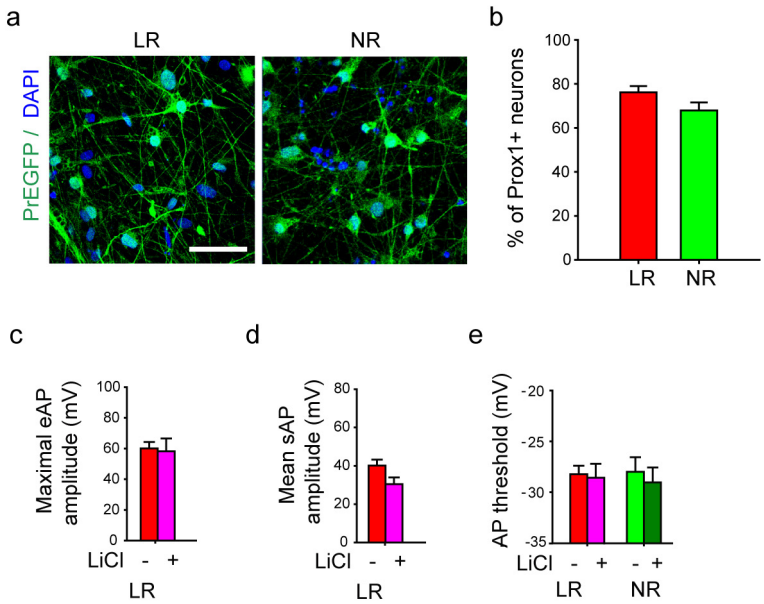
Extended Data Figure 2 | Lentiviral transduction of Prox1::eGFP efficiently labels Prox1-positive DG granule cell-like neurons. a, Sample immunostaining images showing the expression of Prox1 and Prox1::eGFP in the normal and BD neurons. Scale bar, 100 μm. **b**, Bar graph showing that, both in the normal and in BD groups, more than 90% of Prox1::eGFP-positive neurons express nuclear Prox1 protein. Normal, 92.1 ± 2.4%, $n = 4$ lines; BD, 93.3 ± 1.2%, $n = 12$ lines. **c**, Bar graph showing that, both in the normal and in BD groups, approximately 90% of Prox1-positive DG-like neurons express Prox1::eGFP. Bars, mean ± s.e.m.



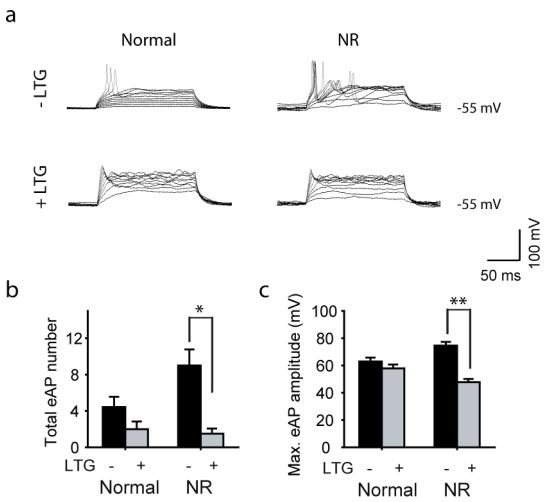
Extended Data Figure 3 | Bar graphs summarizing the similarity between different cell lines of the same subject and comparison of low and high passage cells. a, **b**, Bar graph comparing the MMP ($n = 20$ lines) (**a**) and mitochondria size ($n = 68$ images from 20 lines) (**b**) of different cell lines of one subject. **c**–**f**, Electrophysiological recording experiments, including peak Na^+ currents ($n = 92$ neurons from 20 lines) (**c**), AP threshold (94 neurons from 20 lines) (**d**), total evoked AP number ($n = 97$ neurons from 20 lines) (**e**) and maximal AP amplitude ($n = 96$ neurons from 20 lines) (**f**). **g**, Bar graph comparing the frequency of Ca^{2+} transient events. Black bar, cell line/clone 1; grey bar, cell line/clone 2 (178 videos from 20 lines). **h**, Bar graph showing the normalized peak Na^+ current in normal (NM) and BD neurons derived from <P5 and >P9 cell lines (P5: normal, $n = 40$ neurons from 8 lines; BD, $n = 52$ from 12 lines. P9: normal, $n = 11$ from 2 lines; BD, $n = 23$ from 5 lines). Student's t-test, $^*P < 0.05$. Bars, mean ± s.e.m.



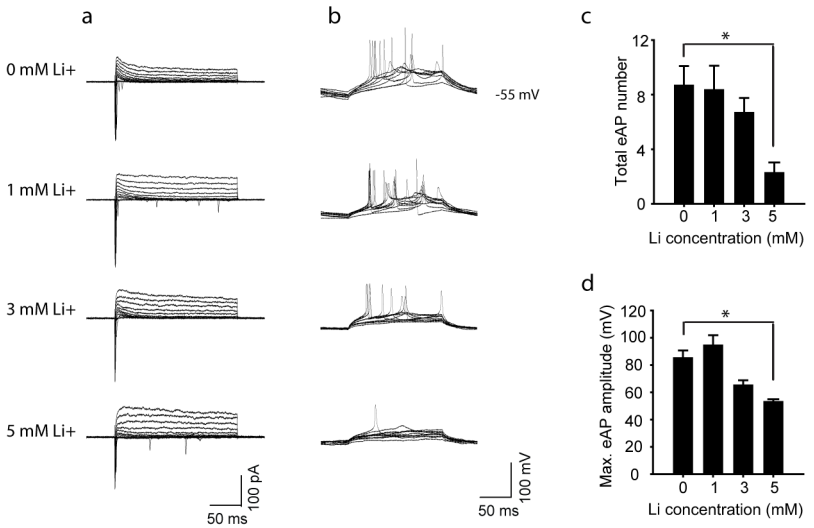
Extended Data Figure 4 | K^+ currents in the BD neurons. a, Average peak values of K^+ currents in the BD and normal neurons. **b**, Normalized average K^+ currents at different membrane potentials (normal, $n = 35$ neurons from 7 lines; BD, $n = 41$ from 10 lines). Student's t-test. Bars, mean ± s.e.m.



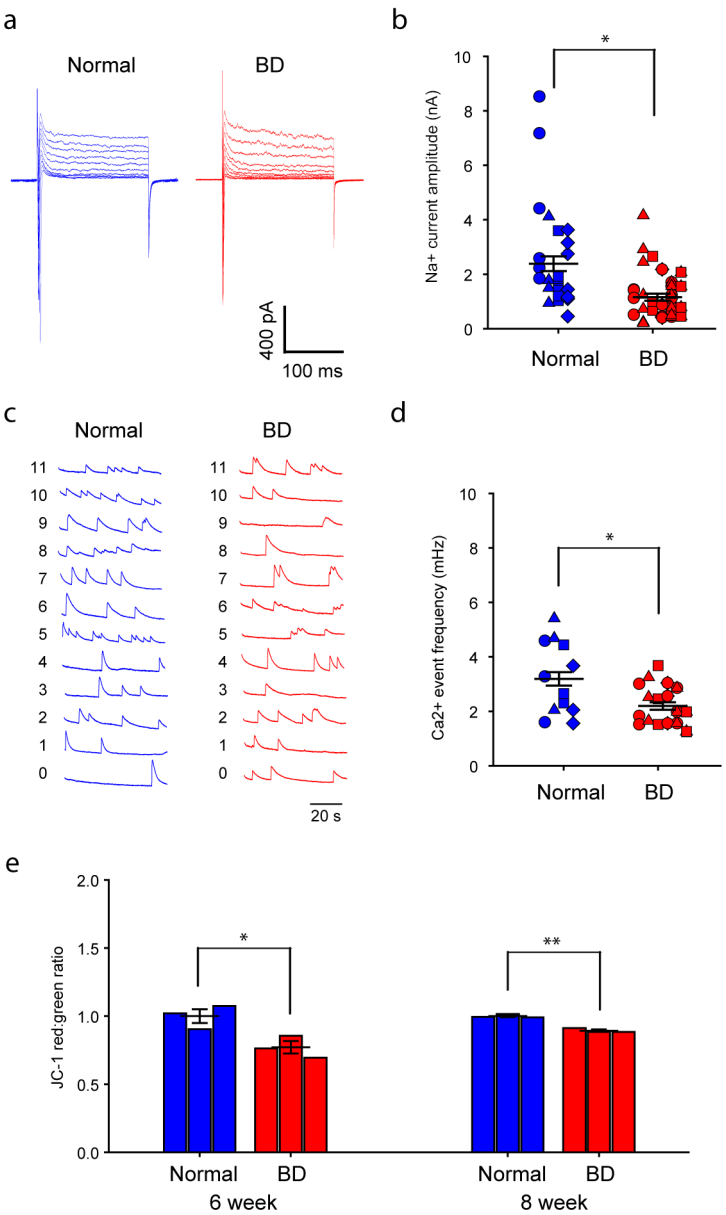
Extended Data Figure 5 | Prox1:eGFP expression in the BD LR and NR neurons and AP amplitude and threshold of BD neurons. a, Sample immunostaining images showing the expression of Prox1::eGFP in the BD LR and NR neurons. Scale bar, 50 µm. **b**, Quantitative analysis revealed a similar percentage of Prox1::eGFP-positive DG-like neurons in the LR and NR groups ($n = 32$ images from 4 lines). **c**, **d**, Bar graphs showing the Li-induced effects in the maximal amplitude of evoked APs (LR without Li treatment, $n = 27$ neurons from 5 lines; with Li treatment, $n = 18$ from 5 lines) (**c**) and mean amplitude of spontaneous APs (LR without Li, $n = 11$ neurons from 3 lines; with Li, $n = 10$ from 3 lines) (**d**) of the LR neurons. **e**, Bar graph showing that the threshold of AP firing was not changed by Li (LR without Li, $n = 11$ neurons from 3 lines; with Li, $n = 10$ from 3 lines). Bars, mean ± s.e.m.



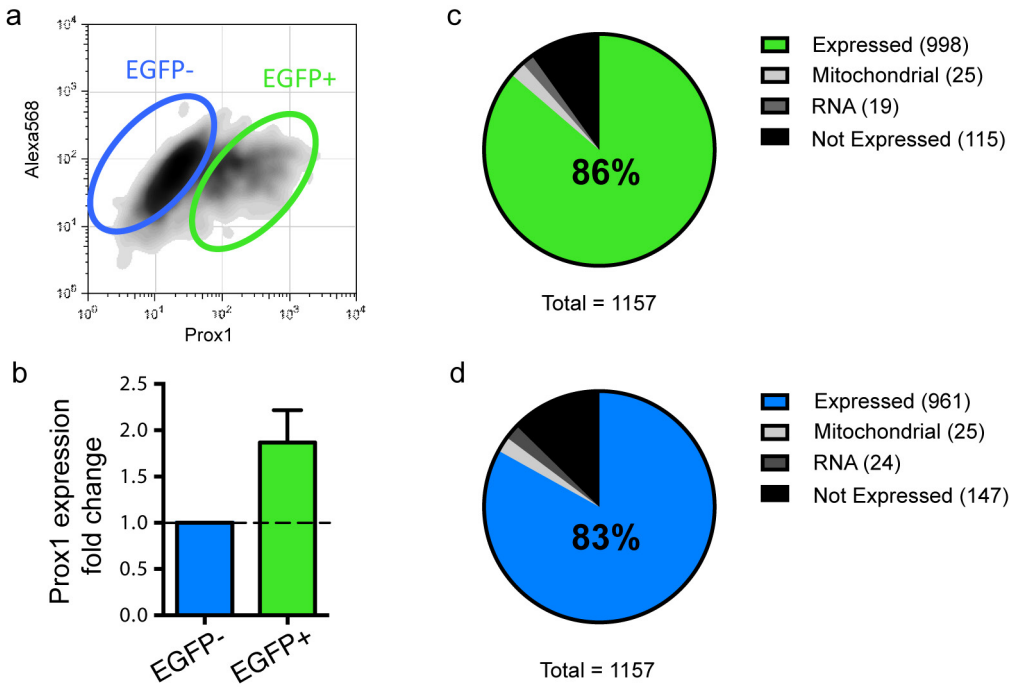
Extended Data Figure 6 | AP firing in the BD NR neurons treated with lamotrigine (LTG). a, Representative traces of APs evoked during 300 ms stepwise depolarization periods in the normal and NR neurons with and without 100 μm lamotrigine treatment. **b**, **c**, Bar graphs summarizing the effects of lamotrigine on the total number (**b**) and maximal amplitude (**c**) of evoked APs in the normal and BD NR neurons (normal: without lamotrigine, $n = 7$ neurons; with lamotrigine, $n = 8$. BD NR: without lamotrigine, $n = 5$; with lamotrigine, $n = 6$). Student's t-test, $^{*}P < 0.05$; $^{**}P < 0.001$. Bars, mean ± s.e.m.



Extended Data Figure 7 | Effect of Li on the normal neurons. **a**, Representative traces of Na^+/K^+ currents in the normal neurons treated with Li at different concentrations. **b**, Representative traces of APs evoked during 300 ms stepwise depolarization periods in the normal neurons treated with Li at different concentrations. **c**, **d**, Bar graphs summarizing the effects of different concentrations of Li on the total number (**c**) and maximal amplitude (**d**) of evoked APs in the normal neurons ($n = 4$ neurons). Student's t-test. $^*P < 0.05$. Bars, mean ± s.e.m.

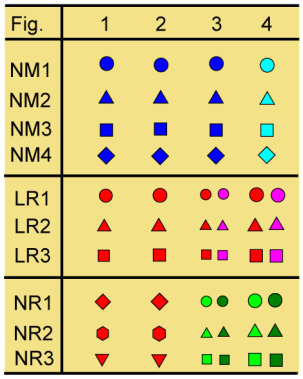


Extended Data Figure 8 | Reversal of hyperexcitability in old BD neurons. **a**, **b**, Sample traces (**a**) and scatter graph (**b**) showing that the 8-week-old BD neurons exhibited weaker Na^+ currents than the normal neurons (normal, $n = 28$ neurons from 4 lines; BD, $n = 37$ from 6 lines). **c**, **d**, Sample traces (**c**) and scatter graph (**d**) showing that the 8-week-old BD neurons exhibited a lower frequency of Ca^{2+} transient events than the normal neurons ($n = 30$ videos from 10 patients). **e**, Scatter graphs showing the MMP of 6- and 8-week-old BD and normal neurons (normal, $n = 3$ lines; BD, $n = 3$ lines). Student's *t*-test, $*P < 0.05$; $**P < 0.001$. Bars, mean ± s.e.m.



Extended Data Figure 9 | The Prox1::eGFP-positive BD cells have similar expression of differentially expressed genes to the whole differentiation culture. a, Sorting of cells strongly expressing Prox1::eGFP using flow cytometry. **b**, Bar graph showing that Prox1::eGFP expression is enriched in the selected cells. **c**, Enrichment of differentially expressed genes in the Prox1 + DG-like neurons (**c**) and non-DG cells (**d**) ($n = 6$ patients). Bars, mean ± s.e.m.

a



Extended Data Figure 10 | Representative icons of the subjects in the figures. a, Representative icons of the patients with BD and healthy people used in the experiments shown in the figures. Identical symbols indicate the same subject.

LETTER

doi:10.1038/nature15376

Microenvironment–induced PTEN loss by exosomal microRNA primes brain metastasis outgrowth

Lin Zhang[1,2]*, Siyuan Zhang[1,3]*, Jun Yao[1], Frank J. Lowery[1,2], Qingling Zhang[1], Wen-Chien Huang[1], Ping Li[1], Min Li[1], Xiao Wang[1], Chenyu Zhang[1], Hai Wang[1], Kenneth Ellis[1], Mujeeburahiman Cheerathodi[4], Joseph H. McCarty[4], Diane Palmieri[5], Jodi Saunus[6], Sunil Lakhani[6,7,8], Suyun Huang[4], Aysegul A. Sahin[9], Kenneth D. Aldape[9], Patricia S. Steeg[5] & Dihua Yu[1,2,10]

The development of life-threatening cancer metastases at distant organs requires disseminated tumour cells' adaptation to, and co-evolution with, the drastically different microenvironments of metastatic sites[1]. Cancer cells of common origin manifest distinct gene expression patterns after metastasizing to different organs[2]. Clearly, the dynamic interaction between metastatic tumour cells and extrinsic signals at individual metastatic organ sites critically effects the subsequent metastatic outgrowth[3,4]. Yet, it is unclear when and how disseminated tumour cells acquire the essential traits from the microenvironment of metastatic organs that prime their subsequent outgrowth. Here we show that both human and mouse tumour cells with normal expression of PTEN, an important tumour suppressor, lose PTEN expression after dissemination to the brain, but not to other organs. The PTEN level in PTEN-loss brain metastatic tumour cells is restored after leaving the brain microenvironment. This brain microenvironment-dependent, reversible PTEN messenger RNA and protein downregulation is epigenetically regulated by microRNAs from brain astrocytes. Mechanistically, astrocyte-derived exosomes mediate an intercellular transfer of PTEN-targeting microRNAs to metastatic tumour cells, while astrocyte-specific depletion of PTEN-targeting microRNAs or blockade of astrocyte exosome secretion rescues the PTEN loss and suppresses brain metastasis *in vivo*. Furthermore, this adaptive PTEN loss in brain metastatic tumour cells leads to an increased secretion of the chemokine CCL2, which recruits IBA1-expressing myeloid cells that reciprocally enhance the outgrowth of brain metastatic tumour cells via enhanced proliferation and reduced apoptosis. Our findings demonstrate a remarkable plasticity of PTEN expression in metastatic tumour cells in response to different organ microenvironments, underpinning an essential role of co-evolution between the metastatic cells and their microenvironment during the adaptive metastatic outgrowth. Our findings signify the dynamic and reciprocal cross-talk between tumour cells and the metastatic niche; importantly, they provide new opportunities for effective anti-metastasis therapies, especially of consequence for brain metastasis patients.

The remarkable phenotypic plasticity observed in metastasis is indicative of co-evolution occurring at specific metastatic organ microenvironments[5,6]. To obtain insights into how disseminated tumour cells acquire essential traits from metastatic microenvironments for successful outgrowth, we analysed public gene expression profiles of clinical metastases from distinct organs as well as organ-specific metastases from mice injected with various cancer cells (Extended Data Fig. 1a–c). Notably, *PTEN* mRNA was markedly downregulated in brain metastases compared to primary tumours or other organ metastases. Our immunohistochemistry (IHC) analyses of PTEN expression confirmed a significantly higher rate of PTEN loss (defined by an immunoreactive score (IRS) of 0–3)[7] in brain metastases (71%) than in unmatched primary breast cancers (30%) (Fig. 1a). PTEN loss was also detected at a significantly higher frequency in brain metastases (71%) than in matched primary breast cancers (37%) of an independent patient cohort (Fig. 1b).

To test a possible role for PTEN loss in brain metastasis[8,9], we intracarotidly injected PTEN-knockdown tumour cells and assessed experimental brain metastasis; unexpectedly, neither incidence nor size of brain metastases was increased (Fig. 1c). Furthermore, patients with PTEN-normal or PTEN-loss primary tumours had comparable levels of brain-metastasis-free survival, and patients with or without brain metastases had similar PTEN levels in their primary tumours (Extended Data Fig. 1d, e). Thus, the observed PTEN loss in brain metastases was unlikely to be derived from PTEN-low primary tumours. To investigate whether PTEN loss in brain metastasis is a secondary non-genetic event imposed by the brain microenvironment, we injected five PTEN-normal breast cancer cell lines either into mammary fat pad (MFP) or intracarotidly to induce brain metastasis. Notably, the PTEN level was significantly decreased in brain metastases compared to the respective MFP tumours or lung metastases (Extended Data Fig. 2a, b). We repeated the injections with cells clonally expanded from single PTEN-normal tumour cells, and observed similar phenotypes (Fig. 1d), suggesting that PTEN-loss brain metastases were not selected from pre-existing PTEN-low cells in the primary tumours. Surprisingly, established sublines from PTEN-low brain metastases (primary Br cells) regained PTEN expression in culture comparable to parental cells (Fig. 1e). Analogously, two *in-vivo*-selected brain-seeking sublines exhibited similar PTEN levels to their matched parental cells *in vitro* (Extended Data Fig. 2c). Re-injecting the cultured PTEN-normal primary brain sublines conferred a distinct PTEN loss in secondary brain metastases, but not in secondary MFP tumours, and PTEN levels in secondary brain subline cells were fully restored again in culture (Fig. 1f, g and Extended Data Fig. 2d), indicating a reversible non-genetic PTEN loss in the brain tumour microenvironment (TME).

To explore how the brain TME regulates PTEN in metastatic cells[10–12], we co-cultured tumour cells with primary glia (>90% astrocytes)[13], cancer-associated fibroblasts (CAFs), or NIH3T3 fibroblasts. Co-culture with glia led to a significant decrease of *PTEN* mRNA and PTEN protein (Fig. 2a, b and Extended Data Fig. 2e, f) in all tumour cells, but did not affect *PTEN* promoter methylation or activity (Extended Data Fig. 2g, h). This prompted us to examine whether glia reduce *PTEN* mRNA stability through microRNAs (miRNAs). Five miRNAs (miR-17, miR-19a, miR-19b, miR-20a and miR-92) in the miR-17~92 cluster were functionally demonstrated to target *PTEN* (refs 14–17), and *Mirc1*$^{tm1.1Tyj}$/*J* mice have a floxed miR-17~92 allele[18]. We

[1]Department of Molecular and Cellular Oncology, The University of Texas M. D. Anderson Cancer Center, Houston, Texas 77030, USA. [2]Cancer Biology Program, Graduate School of Biomedical Sciences, Houston, Texas 77030, USA. [3]Department of Biological Sciences, University of Notre Dame, Notre Dame, Indiana 46556, USA. [4]Department of Neurosurgery, The University of Texas M. D. Anderson Cancer Center, Houston, Texas 77030, USA. [5]Woman's Malignancies Branch, National Cancer Institute, Bethesda, Maryland 20892, USA. [6]The University of Queensland Centre for Clinical Research, Brisbane, Queensland 4029, Australia. [7]The School of Medicine and Pathology Queensland, Brisbane, Queensland 4029, Australia. [8]The Royal Brisbane and Women's Hospital, Brisbane, Queensland 4029, Australia. [9]Department of Pathology, The University of Texas M. D. Anderson Cancer Center, Houston, Texas 77030, USA. [10]Center for Molecular Medicine, China Medical University, Taichung 40402, Taiwan.
*These authors contributed equally to this work.

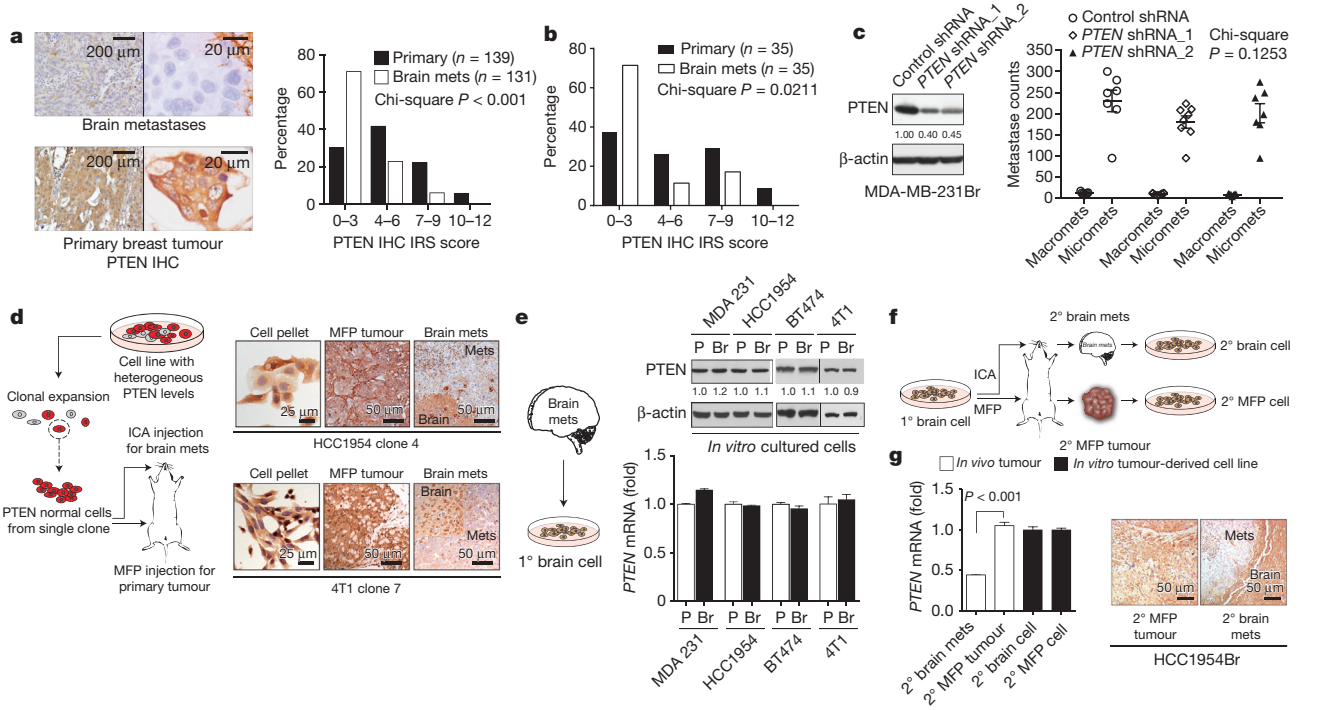


Figure 1 | Brain microenvironment-dependent reversible PTEN down-regulation in brain metastases. **a**, Representative IHC staining and histograms of PTEN protein levels in primary breast tumours ($n = 139$) and unmatched brain metastases (mets) ($n = 131$) (Chi-square test, $P < 0.001$). **b**, Histograms of PTEN protein levels in primary breast tumours and matched brain metastases from 35 patients (Chi-square test, $P = 0.0211$). **c**, PTEN western blots (left) and brain metastasis counts 30 days after intracarotid injection (right) of MDA-MB-231Br cells transfected with control or *PTEN* shRNAs. Macromets: >50 μm in diameter; micromets: ≤50 μm (mean ± s.e.m., Chi-square test, $P = 0.1253$). **d**, PTEN IHC staining of tumours derived from clonally expanded PTEN-normal sublines. ICA, intracarotid artery; MFP, mammary fat pad. **e**, Western blot and quantitative reverse transcriptase PCR (qRT–PCR) of PTEN expression in the indicated parental (P) and brain-seeking (Br) cells under culture (3 biological replicates, with 3 technical replicates each). **f**, Schematic of *in vivo* re-establishment of secondary (2°) brain metastasis, MFP tumour, and their derived cell lines. **g**, *PTEN* qRT–PCR (mean ± s.e.m., *t*-test, 3 biological replicates, with 3 technical replicates each) and PTEN IHC in HCC1954Br secondary tumours and cultured cells.

knocked out the miR-17~92 allele *in situ* in *Mirc1*$^{tm1.1Tyj}$*/J* mice by intracranial injection of astrocyte-specific Cre adenovirus (Ad-GFAP-Cre), then intracarotidly injected syngeneic mouse melanoma B16BL6 cells to form brain metastases (Fig. 2c). Astrocyte-specific depletion of *PTEN*-targeting miRNAs blocked PTEN downregulation (Fig. 2d) in the brain metastasis tumour cells *in vivo* without significantly altering other potential miRNA targets (Extended Data Fig. 3a), and significantly suppressed brain metastasis growth compared to the

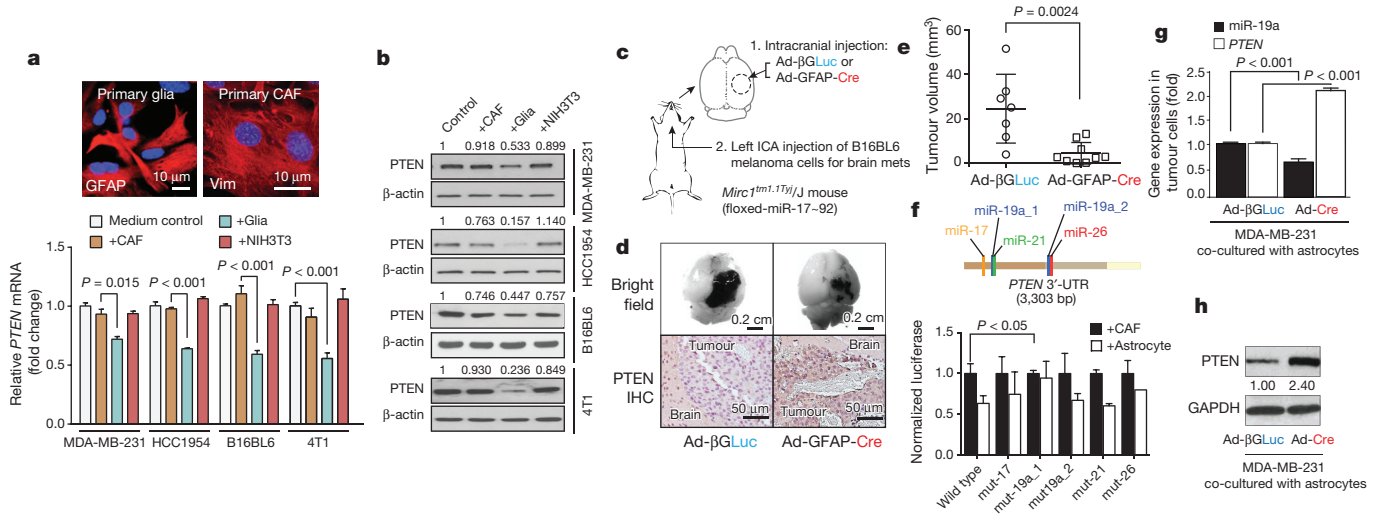


Figure 2 | Astrocyte-derived miRNAs silence PTEN in tumour cells. **a**, *PTEN* mRNA in the indicated tumour cells after 2–5 days co-culture with GFAP-positive primary glia or vimentin (vim)-positive primary CAFs or NIH3T3 fibroblasts (mean ± s.e.m., *t*-test, 3 biological replicates, with 3 technical replicates each). **b**, Western blot of PTEN protein under co-culture as in **a** (3 biological replicates). **c**, Schematic of astrocyte-specific miR-17~92 deletion by GFAP-driven Cre adenovirus (Ad-GFAP-Cre) in *Mirc1*$^{tm1.1Tyj}$*/J* mice. **d**, Representative image of tumour sizes and PTEN IHC of brain metastases. **e**, Quantification of brain metastases volume (mean ± s.d., *t*-test, $P = 0.0024$). **f**, *PTEN* 3′-UTR luciferase activity after co-culture (mean ± s.e.m., *t*-test, 3 biological replicates, with 3 technical replicates each). bp, base pairs. **g**, qRT–PCR analyses of miR-19a and *PTEN* mRNA in MDA-MB-231 cells after 48 h co-culture with primary astrocytes from *Mirc1*$^{tm1.1Tyj}$*/J* mice pre-infected (48 h) by adenovirus (Ad-βGLuc or Ad-GFP-Cre) (mean ± s.e.m., *t*-test, $P < 0.001$, 3 biological replicates, with 3 technical replicates each). **h**, Western blot of PTEN protein in MDA-MB-231 cells, co-cultured as in **g**.

control group (Fig. 2d, e), indicating a tumour cell non-autonomous PTEN downregulation by astrocyte-derived *PTEN*-targeting miRNAs. Astrocyte-specific depletion of *PTEN*-targeting miRNAs also suppressed intracranially injected tumour cell outgrowth (Extended Data Fig. 3b–f). To examine which *PTEN*-targeting miRNA primarily mediates the PTEN loss in tumour cells when co-cultured with astrocytes, the luciferase activities of the wild-type and mutated *PTEN* 3′-untranslated region (UTR) (containing various miRNA binding site mutations) in tumour cells were assessed (Fig. 2f). Compared with CAF co-culture, astrocyte co-culture inhibited luciferase activity of wild-type *PTEN* 3′-UTR, which was rescued by the miR-19a binding site mutation (position 1), but not by other mutations, indicating the major role of miR-19a in astrocyte-mediated *PTEN* mRNA downregulation in tumour cells. Furthermore, *PTEN* mRNA (Fig. 2g and Extended Data Fig. 3g) and PTEN protein (Fig. 2h and Extended Data Fig. 3h) were not downregulated in tumour cells co-cultured with primary astrocytes from *Mirc1*$^{tm1.1Tyj}$/*J* mice in which *PTEN*-targeting miRNAs were depleted (Extended Data Fig. 3i).

After co-culture with Cy3-labelled miR-19a-transfected primary astrocytes, we detected significantly more Cy3$^+$ epithelial cell adhesion molecule (EpCAM)-positive tumour cells over time than under CAF co-culture (Fig. 3a and Extended Data Fig. 4a), suggesting that miR-19a is intercellularly transferred from astrocytes to tumour cells. miRNAs are transferable between neighbouring cells through gap junctions or small vesicles[19,20]. Treating tumour cells with a gap junction channel inhibitor, carbenoxolone disodium salt, had no significant effect on miR-19a intercellular transfer (data not shown), while adding astrocyte-conditioned media to tumour cells led to an increase in miR-19a levels and a subsequent PTEN downregulation (Extended Data Fig. 4b–d). Recognizing the involvement of exosomes in neuronal function and glioma development[21], we postulated that exosomes may mediate miR-19a transfer from astrocytes to tumour cells. Indeed, transmission electron microscopy detected spherical, membrane-encapsulated particles between 30 and 100 nm, typical of exosome vesicles, in astrocyte-conditioned media[22] (Fig. 3b). Additionally, the astrocyte-conditioned media contained significantly more CD63$^+$, CD81$^+$ and TSG101$^+$ exosomes[22] than the CAF-conditioned media (Fig. 3c and Extended Data Fig. 4e, f). Moreover, the exosomes from astrocytes contained 3.5-fold higher levels of miR-19a than those from CAFs (Extended Data Fig. 4g). Adding exosomes purified from conditioned media of Cy3-miR-19a-transfected astrocytes led to miR-19a transfer into cultured tumour cells (Fig. 3d). Furthermore, treating tumour cells directly with astrocyte-derived exosomes led to a dose-dependent increase of miR-19a and a subsequent decrease of *PTEN* mRNA in tumour cells (Fig. 3e). To determine whether astrocyte-released exosomes are required for miR-19a transfer, we blocked astrocyte exosome secretion by treating astrocytes with either an inhibitor of exosome release, dimethyl amiloride (DMA), or a short interfering RNA (siRNA) targeting *Rab27a*, a mediator of exosome secretion[23] (Extended Data Fig. 5a–c). Both exosome blockades

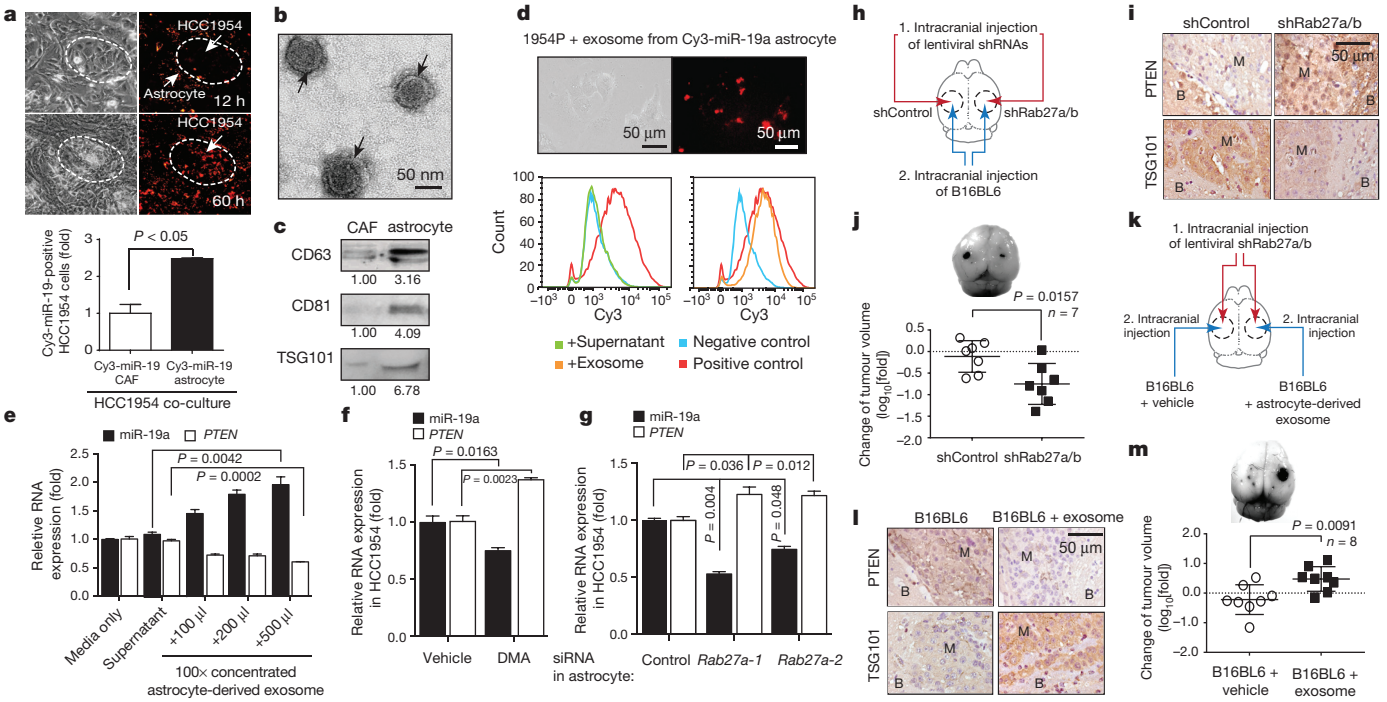


Figure 3 | Intercellular transfer of PTEN-targeting miR-19a to tumour cells via astrocyte-derived exosomes. **a**, Intercellular transfer of miR-19a. Top, light microscopy and fluorescent images of HCC1954 cells 12 and 60 h after co-culture with astrocytes loaded with Cy3-labelled miR-19a. Bottom, flow cytometry analysis of Cy3-miR-19a in tumour cells 60 h after co-culture (mean ± s.e.m., *t*-test, $P < 0.05$, 3 biological replicates). **b**, **c**, Transmission electron microscopy of exosome vesicles in astrocyte-conditioned media (**b**), confirmed by western blot for CD63, CD81 and TSG101 exosome markers released by 1×10^6 CAFs or astrocytes (**c**). **d**, Representative data showing presence of Cy3-miR-19a in HCC1954 breast cancer cells after adding exosomes purified from Cy3-miR-19a-transfected astrocytes for 24 h. Bottom, flow cytometry analysis of Cy3-miR-19a-positive HCC1954 cells after treatment with supernatant (without exosomes), or exosomes purified from Cy3-miR-19a-transfected astrocytes. Negative control is HCC1954 cells without treatment. Positive control is Cy3-miR-19a-transfected astrocytes (3 biological replicates). **e**, Histogram of miR-19a and *PTEN* mRNA in HCC1954 cells 48 h after addition of media, astrocyte supernatant, or exosomes purified from astrocyte-conditioned media (mean ± s.e.m., *t*-test, 3 biological replicates, with 3 technical replicates each). **f**, **g**, Histograms of miR-19a and *PTEN* mRNA in HCC1954 cells after 48 h co-culture in conditioned media from vehicle- or DMA-treated (4 h) astrocytes (**f**) and control- or *Rab27a*-siRNA-transfected (48 h) astrocytes (**g**) (mean ± s.e.m., *t*-test, 3 biological replicates, with 3 technical replicates each). **h**–**j**, Schematics of *in vivo* experiments (**h**), IHC analyses of PTEN and exosome marker expression (**i**) and changes of tumour volume (**j**) (mean ± s.d., *t*-test, $n = 7$, $P = 0.0157$). B, brain; M, metastases; shRab27a/b, shRNA against *Rab27a/b*. **k**–**m**, Schematics showing *in vivo* rescue of exosome effect by pre-incubation of tumour cells with astrocyte-derived exosomes (**k**), IHC analyses of PTEN and exosome marker expression (**l**) and changes of tumour volume (**m**) (mean ± s.d., *t*-test, $n = 8$, $P = 0.0091$).

decreased the transfer of miR-19a from astrocytes to tumour cells and restored the *PTEN* mRNA levels (Fig. 3f, g). Furthermore, we intracranially injected *Rab27a/b* short hairpin RNA (shRNA) lentiviruses to block exosome secretion in mouse brain parenchyma (brain metastasis stroma), and then inoculated B16BL6 melanoma cells to the same sites (Fig. 3h). Inhibiting Rab27a/b reduced $TSG101^+$ and $CD63^+$ exosomes, blocked PTEN downregulation in tumour lesions (Fig. 3i and Extended Data Fig. 5d–g), and significantly decreased tumour outgrowth (Fig. 3j). Conversely, intracranial co-injection of tumour cells with astrocyte-derived exosomes (Fig. 3k) rescued PTEN downregulation in tumour cells (Fig. 3l) and metastatic outgrowth (Fig. 3m) in mouse brains injected with *Rab27a/b* shRNA (Extended Data Fig. 5h, i). Collectively, exosome-mediated miR-19a transfer from astrocytes to tumour cells is critical for tumour PTEN downregulation and aggressive outgrowth in the brain.

We next explored how PTEN loss promotes brain metastasis. Doxycycline-inducible PTEN knockdown (Extended Data Fig. 6a) before intracarotid injection did not alter tumour cell extravasation into the brain parenchyma (Extended Data Fig. 6b, c). To test whether restoring PTEN expression after tumour cell extravasation inhibits metastatic outgrowth, we selected subclones of human breast carcinoma cells that selectively metastasize to the brain (MDA-MB-231Br) stably expressing either a doxycycline-inducible *PTEN*-coding sequence without the 3′-UTR miRNA binding sites, or red fluorescent protein (RFP) controls (Fig. 4a and Extended Data Fig. 6d). PTEN induction 7 days post-intracarotid injection after extravasation of

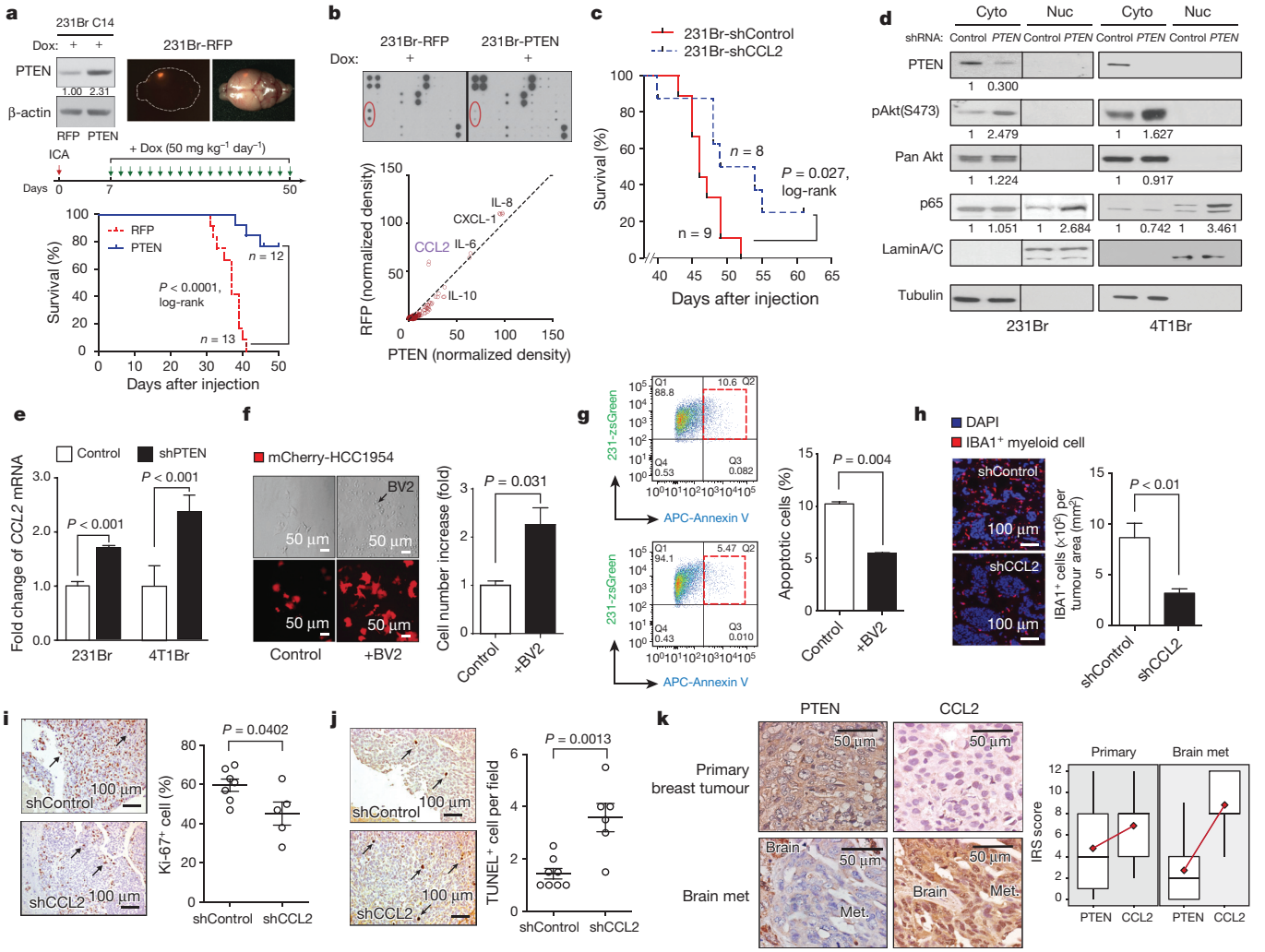


Figure 4 | Brain-dependent PTEN loss instigates metastatic microenvironment to promote metastatic cell outgrowth. **a**, Prolonged mouse survival by restoration of PTEN expression. Top, doxycycline (Dox)-inducible RFP (left) and PTEN expression (right) in 231Br cells. Middle, schematic of brain metastasis assay with doxycycline-induced RFP or PTEN expression. Bottom, overall survival of mice bearing brain metastases of 231Br cells with induced PTEN re-expression or RFP expression (log-rank test, $n = 12$, $P < 0.0001$). **b**, Cytokine array of 231Br cells with doxycycline-induced RFP or PTEN expression. **c**, Overall survival of mice bearing brain metastases of 231Br cells transfected with control or *CCL2* shRNAs (shControl or shCCL2, respectively) (log-rank test, $n = 8$, $P = 0.027$). **d**, Western blot analysis of NF-κB p65 nuclear translocation after knocking down PTEN. Cyto, cytosol; nuc, nuclear. **e**, Histogram showing *CCL2* mRNA levels detected by quantitative PCR after PTEN knockdown with shRNA (shPTEN) (mean ± s.e.m., *t*-test, $P < 0.001$, 3 biological replicates, with 3 technical replicates each). **f**, Light and fluorescent microscopy images and quantification of mCherry-labelled tumour cells with or without BV2 microglia co-culture under 2-day serum starvation (mean ± s.e.m., *t*-test, $P = 0.031$, 3 biological replicates, with 3 technical replicates each). **g**, FACS analyses of Annexin V^+ apoptotic zsGreen-labelled 231Br cells under doxorubicin treatment with or without BV2 microglia co-culture (mean ± s.e.m., *t*-test, $P = 0.004$, 3 biological replicates). **h**, Immunofluorescence staining of $IBA1^+$ myeloid cells in brain metastases of 231Br cells containing control (shControl) or *CCL2* shRNA (shCCL2) (mean ± s.e.m., *t*-test, $P < 0.01$, 3 biological replicates, with 3 technical replicates each). **i**, **j**, IHC analyses showing decreased proliferation (Ki-67, **i**) and increased apoptosis (TUNEL staining, **j**) in brain metastases after shRNA-mediated CCL2 knockdown *in vivo* (mean ± s.e.m., *t*-test). **k**, PTEN and CCL2 expression in matched primary breast tumours and brain metastases. Left, representative IHC staining of PTEN and CCL2. Right, quantification of PTEN and CCL2 expression in 35 cases of matched primary breast tumours and brain metastases (mean ± s.d.).

tumour cells markedly extended the overall survival of brain metastases-bearing mice (Fig. 4a and Extended Data Fig. 6e). Collectively, PTEN loss primes brain metastasis outgrowth after tumour cell extravasation and PTEN restoration suppresses the outgrowth.

Autocrine and paracrine signalling have decisive roles in metastasis seeding and outgrowth. Although PTEN restoration only led to a trend of reduced Akt and P70S6K phosphorylation (pAkt and pP70S6K, respectively; Extended Data Fig. 6f), cytokine array analyses revealed markedly reduced CCL2 secretion in PTEN-expressing tumour cells compared to controls (Fig. 4b); whereas PTEN knockdown increased CCL2 expression (Extended Data Fig. 6g). Moreover, the overall survival of brain metastasis-bearing mice with CCL2-knockdown MDA-MB-231Br cells was significantly extended compared to controls (Fig. 4c and Extended Data Fig. 6h, i). Mechanistically, PTEN induction decreased NF-κB p65 phosphorylation (Extended Data Fig. 7a, b) along with reduced CCL2 secretion (Fig. 4b), whereas PTEN knockdown increased p65 nuclear translocation, an indicator of NF-κB activation, and CCL2 expression (Fig. 4d, e), partly through Akt activation (Extended Data Fig. 7c). Furthermore, *CCL2* mRNA and CCL2 protein expression in brain-seeking tumour cells was inhibited by the NF-κB inhibitor pyrrolidine dithiocarbamate (PDTC) (Extended Data Fig. 7d–f), indicating that NF-κB activation is crucial for PTEN-loss-induced CCL2 upregulation.

CCL2 is a chemo-attractant during inflammation[24]. CCL2 receptor (CCR2)-expressing brain-derived IBA1-positive ($IBA1^+$) primary myeloid cells and BV2 microglial cells (Extended Data Fig. 8a, b) migrate towards CCL2, which was blocked by CCR2 antagonists[25] (Extended Data Fig. 8c, d). Functionally, co-culturing with BV2 cells enhanced proliferation and inhibited apoptosis of breast cancer cells (Fig. 4f, g). *In vivo*, CCL2-knockdown brain metastases had decreased $IBA1^+/CCR2^+$ myeloid cell infiltration (Fig. 4h), corresponding to their reduced proliferation and increased apoptosis (Fig. 4i, j). Furthermore, IHC staining of human primary breast tumours and matched brain metastases for PTEN and CCL2 (Figs 1b and 4k, respectively) revealed a significantly ($P = 0.027$) higher CCL2 expression in brain metastases than in primary tumours (Extended Data Fig. 9a). Importantly, severe PTEN loss in brain metastases corresponded to higher CCL2 expression (Extended Data Fig. 9b), which significantly correlated with $IBA1^+$ myeloid cell recruitment (Extended Data Fig. 9c), validating that PTEN downregulation in brain metastatic tumour cells contributes to CCL2 upregulation and $IBA1^+$ myeloid cell recruitment in clinical brain metastases.

Taken together, our data unveiled a complex reciprocal communication between metastatic tumour cells and their TME, which primes the successful outgrowth of cancer cells to form life-threatening metastases (Extended Data Fig. 10). Beyond a tumour cell autonomous view of metastasis, our findings highlighted an important plastic and tissue-dependent nature of metastatic tumour cells, and a bi-directional co-evolutionary view of the 'seed and soil' hypothesis. Notably, although clinical application of CCL2 inhibitor for metastasis treatment requires careful design[26], our data of brain metastasis inhibition by stable ablation of PTEN-loss-induced CCL2 demonstrated the potential of CCL2-targeting for therapeutic intervention of life-threatening brain metastases.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 1 October 2014; accepted 5 August 2015.
Published online 19 October 2015.

1. Quail, D. F. & Joyce, J. A. Microenvironmental regulation of tumor progression and metastasis. *Nature Med.* **19,** 1423–1437 (2013).
2. Park, E. S. *et al.* Cross-species hybridization of microarrays for studying tumor transcriptome of brain metastasis. *Proc. Natl Acad. Sci. USA* **108,** 17456–17461 (2011).
3. Joyce, J. A. & Pollard, J. W. Microenvironmental regulation of metastasis. *Nature Rev. Cancer* **9,** 239–252 (2009).
4. Vanharanta, S. & Massagué, J. Origins of metastatic traits. *Cancer Cell* **24,** 410–421 (2013).
5. Gray, J. Cancer: genomics of metastasis. *Nature* **464,** 989–990 (2010).
6. Friedl, P. & Alexander, S. Cancer invasion and the microenvironment: plasticity and reciprocity. *Cell* **147,** 992–1009 (2011).
7. Zhang, S. *et al.* Combating trastuzumab resistance by targeting SRC, a common node downstream of multiple resistance pathways. *Nature Med.* **17,** 461–469 (2011).
8. Gonzalez-Angulo, A. M. *et al.* PI3K pathway mutations and PTEN levels in primary and metastatic breast cancer. *Mol. Cancer Ther.* **10,** 1093–1101 (2011).
9. Wikman, H. *et al.* Relevance of PTEN loss in brain metastasis formation of breast cancer patients. *Breast Cancer Res.* **14,** R49 (2012).
10. Song, M. S., Salmena, L. & Pandolfi, P. P. The functions and regulation of the PTEN tumour suppressor. *Nature Rev. Mol. Cell Biol.* **13,** 283–296 (2012).
11. Hopkins, B. D. *et al.* A secreted PTEN phosphatase that enters cells to alter signaling and survival. *Science* **341,** 399–402 (2013).
12. Miething, C. *et al.* PTEN action in leukaemia dictated by the tissue microenvironment. *Nature* **510,** 402–406 (2014).
13. Mecha, M. *et al.* An easy and fast way to obtain a high number of glial cells from rat cerebral tissue: A beginners approach. *Protocol Exchange* http://dx.doi.org/10.1038/protex.2011.218 (2011).
14. Zhu, H., Han, C., Lu, D. & Wu, T. miR-17–92 cluster promotes cholangiocarcinoma growth: evidence for PTEN as downstream target and IL-6/Stat3 as upstream activator. *Am. J. Pathol.* **184,** 2828–2839 (2014).
15. Liu, S.-Q., Jiang, S., Li, C., Zhang, B. & Li, Q.-J. miR-17–92 cluster targets phosphatase and tensin homology and Ikaros Family Zinc Finger 4 to promote TH17-mediated inflammation. *J. Biol. Chem.* **289,** 12446–12456 (2014).
16. Olive, V. *et al.* miR-19 is a key oncogenic component of mir-17–92. *Genes Dev.* **23,** 2839–2849 (2009).
17. Olive, V., Jiang, I. & He, L. mir-17-92, a cluster of miRNAs in the midst of the cancer network. *Int. J. Biochem. Cell Biol.* **42,** 1348–1354 (2010).
18. Ventura, A. *et al.* Targeted deletion reveals essential and overlapping functions of the miR-17 through 92 family of miRNA clusters. *Cell* **132,** 875–886 (2008).
19. Mittelbrunn, M. *et al.* Unidirectional transfer of microRNA-loaded exosomes from T cells to antigen-presenting cells. *Nature Commun.* **2,** 282 (2011).
20. Suetsugu, A. *et al.* Imaging exosome transfer from breast cancer cells to stroma at metastatic sites in orthotopic nude-mouse models. *Adv. Drug Deliv. Rev.* **65,** 383–390 (2013).
21. Frühbeis, C., Fröhlich, D. & Krämer-Albers, E. M. Emerging roles of exosomes in neuron–glia communication. *Front. Physiol.* **3,** 119 (2012).
22. Kesimer, M. *et al.* Characterization of exosome-like vesicles released from human tracheobronchial ciliated epithelium: a possible role in innate defense. *FASEB J.* **23,** 1858–1868 (2009).
23. Peinado, H. *et al.* Melanoma exosomes educate bone marrow progenitor cells toward a pro-metastatic phenotype through MET. *Nature Med.* **18,** 883–891 (2012).
24. Conti, I. & Rollins, B. J. CCL2 (monocyte chemoattractant protein-1) and cancer. *Semin. Cancer Biol.* **14,** 149–154 (2004).
25. Qian, B.-Z. *et al.* CCL2 recruits inflammatory monocytes to facilitate breast-tumour metastasis. *Nature* **475,** 222–225 (2011).
26. Bonapace, L. *et al.* Cessation of CCL2 inhibition accelerates breast cancer metastasis by promoting angiogenesis. *Nature* **515,** 130–133 (2014).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank M.-C. Hung, H.-K. Lin and Z. Lu for reading the manuscript, A. Yung for *PTEN* promoter constructs, MD Anderson Cancer Center (MDACC) shRNA and ORFome, FACS, histology and high-resolution electron microscopy support, the animal core facilities (NIH CA16672) for technical support, and members of the Yu laboratory for helpful discussions. Thanks to A. Matsika for histological review and tissue microarray construction. This work was supported partially by DOD Center of Excellence grant (P.S.S.) subproject W81XWH-06-2-0033 (D.Y.), NIH Pathway to Independence Award 5R00CA158066-05 (S.Z.), DOD Postdoctoral Fellowship W81XWH-11-1-0003 (C.Z.), Isaiah Fidler Fellowship in Cancer Metastasis (F.J.L.), PO1-CA099031 project 4 (D.Y.), RO1-CA112567-06 (D.Y.), R01CA184836 (D.Y.), Susan G. Komen Breast Cancer Foundation Promise Grant KG091020 (D.Y.), METAvivor Research Grant (D.Y.), Breast and Ovarian Cancers Moon Shot program, China Medical University Research Fund, and Sowell-Huggins Pre-doctoral Fellowship (L.Z.) and Professorship (D.Y.) in Cancer Research. D.Y. is the Hubert L. & Olive Stringer Distinguished Chair in Basic Science at the MDACC.

Author Contributions L.Z., S.Z. and D.Y. developed original hypothesis and designed experiments. L.Z., S.Z., J.Y., F.J.L., Q.Z., W.-C.H., P.L., M.L., X.W., C.Z., K.E., H.W., D.P., M.C., J.H.M., P.S.S. and D.Y. performed experiments and/or analysed data. J.S., S.L., S.H., A.A.S., K.D.A. and P.S.S. provided critical reagents and/or clinical samples. S.Z., L.Z., F.J. and D.Y. wrote and edited the manuscript. D.Y. supervised the study.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to D.Y. (dyu@mdanderson.org).

METHODS

Reagents and cell culture. All common chemicals were from Sigma. Pyrrolidinedithiocarbamic acid was from Santa Cruz Biotechnology. Exo-FBS exosome-depleted FBS was purchased from System Biosciences (SBI). PTEN (9188), pAkt(T308) (9275), pAkt(S473) (4060), Pan Akt (4691), and Bim (2933) antibodies were from Cell Signaling. CD9 (ab92726), Rab27a (ab55667), AMPK (ab3759), CCL2 (ab9899), MAP2 (ab11267), and pP70S6K (ab60948) antibodies were from Abcam. Tsg101 (14497-1-AP) and Rab27b (13412-1-AP) antibodies were from Proteintech. CD81 (104901) antibody was from BioLegend. E2F1 (NB600-210) and CCR2 (NBP1-48338) antibodies were from Novus. GFAP (Z0334) antibody was from DAKO. IBA1 antibody was from WAKO. Cre (969050) antibody was from Novagen. NF-κB p65 (SC-109) and CD63 (SC-15363) antibodies were from Santa Cruz. DMA (sc-202459) and CCR2 antagonist (sc-202525) were from Santa Cruz. MK2206 (S1078) was from Selleckchem. PDTC (P8765) was from Sigma-Aldrich. Human breast cancer cell lines (MDA-MB-231, HCC1954, BT474 and MDA-MB-435) and mouse cell lines (B16BL6 mouse melanoma and 4T1 mouse breast cancer) were purchased from ATCC and verified by the MD Anderson Cancer Center (MDACC) Cell Line Characterization Core Facility. All cell lines have been tested for mycoplasma contamination. Primary glia was isolated as described[13]. In brief, after homogenization of dissected brain from postnatal day (P)0–P2 neonatal mouse pups, all cells were seeded on poly-D-lysine coated flasks. After 7 days, flasks with primary culture were placed on an orbital shaker and shaken at 230 r.p.m. for 3 h. Warm DMEM 10:10:1 (10% of fetal bovine serum, 10% of horse serum, 1% penicillin/streptomycin) was added and flasks were shaken again at 260 r.p.m. overnight. After shaking, fresh trypsin was added into the flask and leftover cells were plated with warm DMEM 5:5:1 (5% of fetal bovine serum, 5% of horse serum, 1% penicillin/streptomycin) to establish primary astrocyte culture. More than 90% of isolated primary glial cells were $GFAP^+$ astrocytes. Primary CAFs were isolated by digesting the mammary tumours from MMTV-neu transgenic mouse. 231-xenograft CAFs were isolated by digesting the mammary tumours from MDA-MB-231 xenograft. For the mixed co-culture experiments, tumour cells were mixed with an equal number of freshly isolated primary glia, CAFs or NIH3T3 fibroblast cells in six-well plate (1:3 ratio). Co-cultures were maintained for 2–5 days before magnetic-bead-based separation. For the trans-well co-culture experiments, tumour cells were seeded in the bottom well and freshly isolated primary glia, CAFs or NIH3T3 cells were seeded on the upper insert (1:3 ratio). Co-cultures were maintained for 2–5 days for the further experiments. Lentiviral-based packaging vectors (Addgene), pLKO.1 PTEN-targeting shRNAs and all siRNAs (Sigma), Human Cytokine Antibody Array 3 (Ray biotech), and lentiviral-based vector pTRIPZ-PTEN and pTRIPZ-CCL2 shRNAs (MDACC shRNA and ORFome Core, from Open Biosystems) were purchased. The human PTEN-targeting shRNA sequences in the lentiviral constructs were: 5′-CCGGAGGCGCTATGTGTATTATT ATCTCGAGATAATAATACACATAGCGCCTTTTTT-3′ (targeting coding sequence); 5′-CCGGCCACAAATGAAGGGATATAAACTCGAGTTTATAT CCCTTCATTTGTGGTTTTT-3′ (targeting 3′-UTR). The human PTEN-targeting siRNA sequences used were: 5′-GGUGUAAUGAUAUGUGCAU-3′ and 5′-GUUAAAGAAUCAUCUGGAU-3′. The human CCL2-targeting siRNA sequences used were: 5′-CAGCAAGUGUCCCAAAGAA-3′ and 5′-CCGAAGA CUUGAACACUCA-3′. The mouse *Rab27a*-targeting siRNA sequences used were: 5′-CGAUUGAGAUGCUCCUGGA-3′ and 5′-GUCAUUUAGGGAUCC AAGA-3′. Mouse pLKO shRNA (shRab27a: TRCN0000381753; shRab27b: TRCN0000100429) were purchased from Sigma. For lentiviral production, lentiviral expression vector was co-transfected with the third-generation lentivirus packing vectors into 293T cells using Lipo293 DNA *in vitro* Transfection Reagent (SignaGen). Then, 48–72 h after transfection, cancer cell lines were stably infected with viral particles. Transient transfection with siRNA was performed using pepMute siRNA transfection reagent (SignaGen). For *in vivo* intracranial virus injection, lentivirus was collected from 15 cm plates 48 h after transfection of packaging vectors. After passing a 0.45 μm filter, all viruses were centrifuged at 25,000 r.p.m (111,000*g*) for 90 min at 4 °C. Viral pellet was suspended in PBS (~200-fold concentrated). The final virus titre ($\sim 1 \times 10^9$ UT ml^{-1}) was confirmed by limiting dilution.

Isolation of tumour cells from co-culture. Cell isolation was performed based on the magnetic bead-based cell sorting protocol according to manufacturer's recommendation (Miltenyi Biotec Inc.). After preparation of a single-cell suspension, tumour cells (HCC1954 or BT474) were stained with primary EpCAM-FITC antibody (130-098-113) (50 μl per 10^7 total cells) and incubated for 30 min in the dark at 4 °C. After washing, the cell pellet was re-suspended and anti-FITC microbeads (50 μl per 10^7 total cells) were added before loading onto the magnetic column of a MACS separator. The column was washed twice and removed from the separator. The magnetically captured cells were flushed out immediately by firmly applying the plunger. The isolated and labelled cells were analysed on a Gallios flow cytometer (Beckman Coulter). For EpCAM-negative MDA-MB-231 tumour cells, FACS sorting (ARIAII, Becton Dickinson) was used to isolate green fluorescent protein $(GFP)^+$ tumour cells from glia or CAFs.

Isolation of $CD11b^+$ cells from mouse primary glia. Isolation of primary glia was achieved by homogenization of dissected brain from P0–P2 mouse pups. After 7 days, trypsin was added and cells were collected. After centrifugation and re-suspension of cell pellet to a single-cell suspension, cells were incubated with $CD11b^+$ microbeads (Miltenyl Biotec) (50 μl per 10^7 total cells) for 30 min at 4 °C. The cells were washed with buffer and $CD11b^+$ cells were isolated by MACS Column. $CD11b^+$ cells were analysed by flow cytometry and immunofluorescence staining.

Western blotting. Western blotting was done as previously described. In brief, cells were lysed in lysis buffer (20 mM Tris, pH 7.0, 1% Triton X-100, 0.5% NP-40, 250 mM NaCl, 3 mM EDTA and protease inhibitor cocktail). Proteins were separated by SDS–PAGE and transferred onto a nitrocellulose membrane. After membranes were blocked with 5% milk for 30 min, they were probed with various primary antibodies overnight at 4 °C, followed by incubation with secondary antibodies for 1 h at room temperature, and visualized with enhanced chemiluminescence reagent (Thermo Scientific).

qRT–PCR. In brief, total RNA was isolated using miRNeasy Mini Kit (Qiagen) and then reverse transcribed using reverse transcriptase kits (iScript cDNA synthesis Kit, Bio-rad). SYBR-based qRT–PCR was performed using pre-designed primers (Life Technologies). miRNA assay was conducted using Taqman miRNA assay kit (Life Technologies). For quantification of gene expression, real-time PCR was conducted using Kapa Probe Fast Universal qPCR, and SYBR Fast Universal qPCR Master Mix (Kapa Biosystems) on a StepOnePlus real-time PCR system (Applied Biosystems). The relative expression of mRNAs was quantified by $2^{-\Delta\Delta Ct}$ with logarithm transformation. Primers used in qRT–PCR analyses are: mouse *Ccl2*: forward, 5′-GTTGGCTCAGCCAGATGCA-3′; reverse: 5′-AGCCT ACTCATTGGGATCATCTTG-3′. Mouse *Actb*: forward: 5′-AGTGTGACGT TGACATCCGT3′; reverse: 5′-TGCTAGGAGCCAGAGCAGTA-3′. Mouse *Pten*: forward: 5′-AACTTGCAATCCTCAGTTTG-3′; reverse: 5′-CTACTTTGATATC ACCACACAC-3′. Mouse *Ccr2* primer: Cat: 4351372 ID: Mm04207877_m1 (Life technologies)

miRNA labelling and transfection. Synthetic miRNAs were purchased from Sigma and labelled with Cy3 by Silencer siRNA labelling kit (Life Technologies). In brief, miRNAs were incubated with labelling reagent for 1 h at 37 °C in the dark, and then labelled miRNAs were precipitated by ethanol. Labelled miRNAs (100 pmoles) were transfected into astrocytes or CAFs in a 10-cm plate. After 48 h, astrocytes and CAFs containing Cy3-miRNAs were co-cultured with tumour cells (at 5:1 ratio).

***PTEN* promoter methylation analysis and luciferase reporter assay of *PTEN* promoter activity.** Genomic DNA was isolated by PreLink genomic DNA mini Kit (Invitrogen), bisulfite conversion was performed by EpiTect Bisulphite Kit and followed by EpiTect methylation-specific PCR (Qiagen). Primers for PTEN CpG island are 5′-TGTAAAACGACGGCCAGTTTGTTATTATTTTTAGGGTTGG GAA-3′ and 5′-CAGGAAACAGCTATGACCCTAAACCTACTTCTCCTCAA CAACC-3′. Luciferase reporter assays were done as previously described[27]. The wild-type *PTEN* promoter driven pGL3-luciferase reporter was a gift from A. Yung. The pGL3-*PTEN* reporter and a control *Renilla* luciferase vector were co-transfected into tumour cells by Lipofectamine 2000 (Life Technologies). After 48 h, tumour cells were co-cultured with astrocytes or CAFs. Another 48 h later, luciferase activities were measured by Dual-Luciferase Report Assay Kit (Promega) on Luminometer 20/20 (Turner Biosystems). The *PTEN* 3′-UTRs with various miRNA binding-site mutations were generated by standard PCR-mediated mutagenesis method and inserted downstream of luciferase reporter gene in pGL3 vector. The activities of the luciferase reporter with the wild-type and mutated *PTEN* 3′-UTRs were assayed as described above.

Exosome isolation and purification. Astrocytes or CAFs were cultured for 48–72 h and exosomes were collected from their culture media after sequential ultracentrifugation as described previously. In brief, cells were collected, centrifuged at 300*g* for 10 min, and the supernatants were collected for centrifugation at 2,000*g* for 10 min, 10,000*g* for 30 min. The pellet was washed once with PBS and purified by centrifugation at 100,000*g* for 70 min. The final pellet containing exosomes was re-suspended in PBS and used for (1) transmission electron microscopy by fixing exosomes with 2% glutaraldehyde in 0.1 M phosophate buffer, pH 7.4; (2) measure of total exosome protein content using BCA Protein Assay normalized by equal number of primary astrocytes and CAF cells; (3) western blotting of exosome marker protein CD63, CD81 and Tsg101; and (4) qRT–PCR by extracting miRNAs with miRNeasy Mini Kit (Qiagen).

Transmission electron microscopy. Fixed samples were placed on 100-mesh carbon-coated, formvar-coated nickel grids treated with poly-L-lysine for about 30 min. After washing the samples on several drops of PBS, samples were

incubated on drops of buffered 1% gluteraldehyde for 5 min, and then washed several times on drops of distilled water. Afterwards, samples were negatively stained on drops of millipore-filtered aqueous 4% uranyl acetate for 5 min. Stain was blotted dry from the grids with filter paper and samples were allowed to dry. Samples were then examined in a JEM 1010 transmission electron microscope (JEOL) at an accelerating voltage of 80 Kv. Digital images were obtained using the AMT Imaging System (Advanced Microscopy Techniques Corp.).

Flow cytometry analysis of exosome marker proteins, Annexin V and CCR2. For exosome detection, 100 µl exosomes isolated from 10-ml conditioned media of astrocytes or CAFs were incubated with 10 µl of aldehyde/sulfate latex beads (4 µm diameter, Life Technologies) for 15 min at 4 °C. After 15 min, PBS was added to make sample volume up to 400 µl, which was incubated overnight at 4 °C under gentle agitation. Exosome-coated beads were washed twice in FACS washing buffer (1% BSA and 0.1% NaN_3 in PBS), and re-suspended in 400 µl FACS washing buffer, stained with 4 µg of phycoerythrin (PE)-conjugated anti-mouse CD63 antibody (BioLegend) or mouse IgG (Santa Cruz Biotechnology) for 3 h at 4 °C under gentle agitation and analysed on a FACS Canto II flow cytometer. Samples were gated on bead singlets based on FCS and SSC characteristics (4 µm diameter). For Annexin V apoptosis assay, after 24 h doxorubicin (2 µM) treatment, the cells were collected, labelled by APC-Annexin V antibody (Biolegend) and analysed on a FACS Canto II flow cytometer. $CD11b^+$ and BV2 cells were stained with CCR2 antibody (Novus) at 4 °C overnight; they were then washed and stained with Alexa Fluor 488 anti-rabbit IgG (Life Technologies) at room temperature for 1 h. The cells were then analysed on a FACS Canto II flow cytometer.

***In vivo* experiments.** All animal experiments and terminal endpoints were carried out in accordance with approved protocols from the Institutional Animal Care and Use Committee of the MDACC. Animal numbers of each group were calculated by power analysis and animals are grouped randomly for each experiment. No blinding of experiment groups was conducted. MFP tumours were established by injection of 5×10^6 tumour cells in 100 µl of PBS:Matrigel mixture (1:1 ratio) orthotopically into the MFP of 8-week-old Swiss nude mice as done previously[28]. Brain metastasis tumours were established by ICA injection of tumour cells (250,000 cells in 0.1 ml HBSS for MDA-MB-231, HCC1954, MDA-MB-435, 4T1 and B16BL6, and 500,000 cells in 0.1 ml HBSS for BT474.m1 into the right common carotid artery as done previously[29]). Mice (6–8 weeks) were randomly grouped into designated groups. Female mice are used for breast cancer experiments, both female and male are used for melanoma experiments. Since the brain metastasis model does not result in visible tumour burdens in living animal, the endpoints of *in vivo* metastasis experiments are based on the presence of clinical signs of brain metastasis, including but not limited to, primary central nervous system disturbances, weight loss, and behavioural abnormalities. Animals are culled after showing the above signs or 1–2 weeks after surgery based on specific experimental designs. Brain metastasis lesions are enumerated as experimental readout. Brain metastases were counted as micromets and macromets. The definition of micromets and macromets are based on a comprehensive mouse and human comparison study previously published[30]. In brief, ten haematoxylin and eosin (H&E)-stained serial sagittal sections (300 µm per section) through the left hemisphere of the brain were analysed for the presence of metastatic lesions. We counted micrometastases (that is, those ≤ 50 µm in diameter) to a maximum of 300 micrometastases per section, and every large metastasis (that is, those > 50 µm in diameter) in each section. Brain-seeking cells from overt metastases and whole brains were dissected and disaggregated in DMEM/F-12 medium using Tenbroeck homogenizer briefly. Dissociated cell mixtures were plated on tissue culture dish. Two weeks later, tumours cells recovered from brain tissue were collected and expanded as brain-seeking sublines (Br.1). For the astrocyte miR-19 knockout mouse model, $Mirc1^{tm1.1Tyj}/J$ mice (Jax lab) (6–8 weeks) were intracranially injected with Ad5-GFAP-Cre virus (Iowa University, Gene Transfer Vector Core) 2 µl (MOI $\sim 10^8$ U μl^{-1}) per point, total four points at the right hemisphere ($n = 9$). Control group ($n = 7$) was injected with the same dose Ad5-RSV-βGLuc (Ad-βGLuc) at the right hemisphere. All intracranial injections were performed by an implantable guide-screw system. One week after virus injection, mice were intracarotidly injected with 2×10^5 B16BL6 tumour cells. After two weeks, whole brains were dissected and fixed in 4% formaldehyde, and embedded in paraffin. Tumour formation, histological phenotypes of H&E-stained sections, and IHC staining were evaluated. Only parenchymal lesions, which are in close proximity of adenovirus injection, were included in our evaluation. Tumour size was calculated as (longest diameter) × (shortest diameter)^2/2. For the intracranial tumour model, $Mirc1^{tm1.1Tyj}/J$ mice (Jax lab) (6–8 weeks) were intracranially injected as described above. Seven mice were used in the experiment. One week later, these mice were intracranially injected with 2.5×10^5 B16BL6 tumour cells at both sides where adenoviruses were injected. After another week, whole brains were dissected and fixed in 4% formaldehyde, and embedded in paraffin. Tumour formation and phenotype were analysed as above.

For the *Rab27a/b* knockdown mouse model, seven C57BL6 mice (Jax lab) (6–8 weeks) were intracranially injected with concentrated lentivirus containing shRab27a and shRab27b (ratio 1:2) 2 µl per point, total three points at the right hemisphere; concentrated control lentivirus containing pLKO.1 scramble were injected at the left hemisphere. All intracranial injections were performed by an implantable guide-screw system. One week later, mice were intracranially injected with 5×10^4 B16BL6 tumour cells at both sides where they had been infected. After one week, whole brains were dissected and fixed in 4% formaldehyde, and embedded in paraffin. Tumour formation, histological phenotypes of H&E-stained sections, IHC staining were evaluated. When performing metastases size quantification, only parenchymal lesions that were in close proximity to the adenovirus injection sites were included in the analyses. Tumour size was calculated as (longest diameter) × (shortest diameter)^2/2. For exosome rescue experiments, eight C57BL6 mice (Jax lab) (6–8 weeks) were intracranially injected with concentrated lentivirus containing shRab27a and shRab27b (ratio 1:2) 2 µl per point, total 3 points at both hemispheres. One week later, these mice were intracranially injected with 5×10^4 B16BL6 tumour cells with 10 µg exosome isolated from astrocyte media at the right sides where they had been injected with lentivirus; 5×10^4 B16BL6 tumour cells with vehicle were injected at the left sides where lentivirus had been injected. After another week, whole brains were dissected and fixed in 4% formaldehyde, and embedded in paraffin. Tumour formation and phenotype were analysed as above.

For *in vivo* extravasation assay, equal numbers of cells labelled with GFP-control shRNA and RFP-PTEN shRNA (Open Biosystems) were mixed and ICA injected. After cardiac perfusion, brains were collected and sectioned through coronal plan on a vibrotome (Leica) into 50-µm slices. Fluorescent cells were then counted. For inducible PTEN expression *in vivo*, mice were given doxycycline (10 µg kg^{-1}) every other day. To quantify brain metastasis incidence and tumour size, brains were excised for imaging and histological examination at the end of experiments. Ten serial sagittal sections every 300 µm throughout the brain were analysed by at least two pathologists who were blinded to animal groups in all above analyses.

Reverse-phase protein array. Reverse-phase protein array of PTEN-overexpressing cells was performed in the MDACC Functional Proteomics core facility. In brief, cellular proteins were denatured by 1% SDS, serial diluted and spotted on nitrocellulose-coated slides. Each slide was probed with a validated primary antibody plus a biotin-conjugated secondary antibody. The signal obtained was amplified using a Dako Cytomation-catalysed system and visualized by DAB colorimetric reaction. The slides were analysed using customized Microvigene software (VigeneTech Inc.). Each dilution curve was fitted with a logistic model ('Super curve fitting' developed at the MDACC) and normalized by median polish. Differential intensity of normalized log values of each antibody between RFP (control) and PTEN-overexpressed cells were compared in GenePattern (http://genepattern.broadinstitute.org). Antibodies with differential expression ($P < 0.2$) were selected for clustering and heat-map analysis. The data clustering was performed using GenePattern.

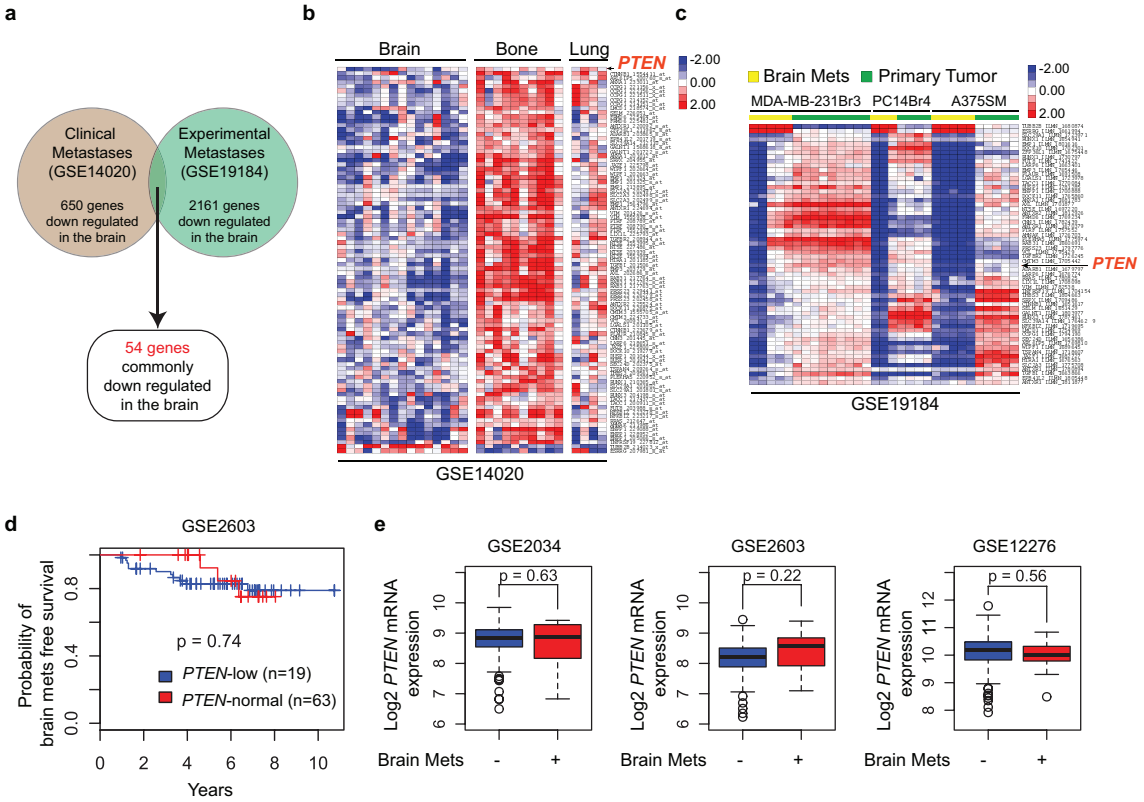
Patient samples. Two studies in separate cohorts were conducted. The first one was a retrospective evaluation of PTEN in two cohorts. (1) Archived formalin-fixed and paraffin-embedded brain metastasis specimens ($n = 131$) from patients with a history of breast cancer who presented with metastasis to the brain parenchyma and had surgery at the MDACC (Supplementary Information). Tissues were collected under a protocol (LAB 02-486) approved by the Institutional Review Board (IRB) at the MDACC. (2) Archived unpaired primary breast cancer formalin-fixed and paraffin-embedded specimens ($n = 139$) collected under an IRB protocol (LAB 02-312) at the MDACC (Supplementary information). Formal consent was obtained from all patients. The second study was a retrospective evaluation of PTEN, CCL2 and IBA1 in the matched primary breast tumours and brain metastatic samples from 35 patients, of which there are 12 HER2-positive, 14 triple-negative and nine oestrogen-receptor-positive tumours according to clinical diagnostic criteria (Supplementary Information). Formalin-fixed, paraffin-embedded primary breast and metastatic brain tumour samples were obtained from the Pathology Department, University of Queensland Centre for Clinical Research. Tissues were collected with approval by human research ethics committees at the Royal Brisbane and Women's Hospital (2005/022) and the University of Queensland (2005000785). For tissue microarray construction, tumour-rich regions (guided by histological review) from each case were sampled using 1-mm cores. All the archival paraffin-embedded tumour samples were coded with no patient identifiers.

IHC and immunofluorescence. Standard IHC staining was performed as described previously[28]. In brief, after de-paraffinization and rehydration, 4 µm sections were subjected to heat-induced epitope retrieval (0.01 M citrate for PTEN). Slides were then incubated with various primary antibodies at 4 °C overnight, after

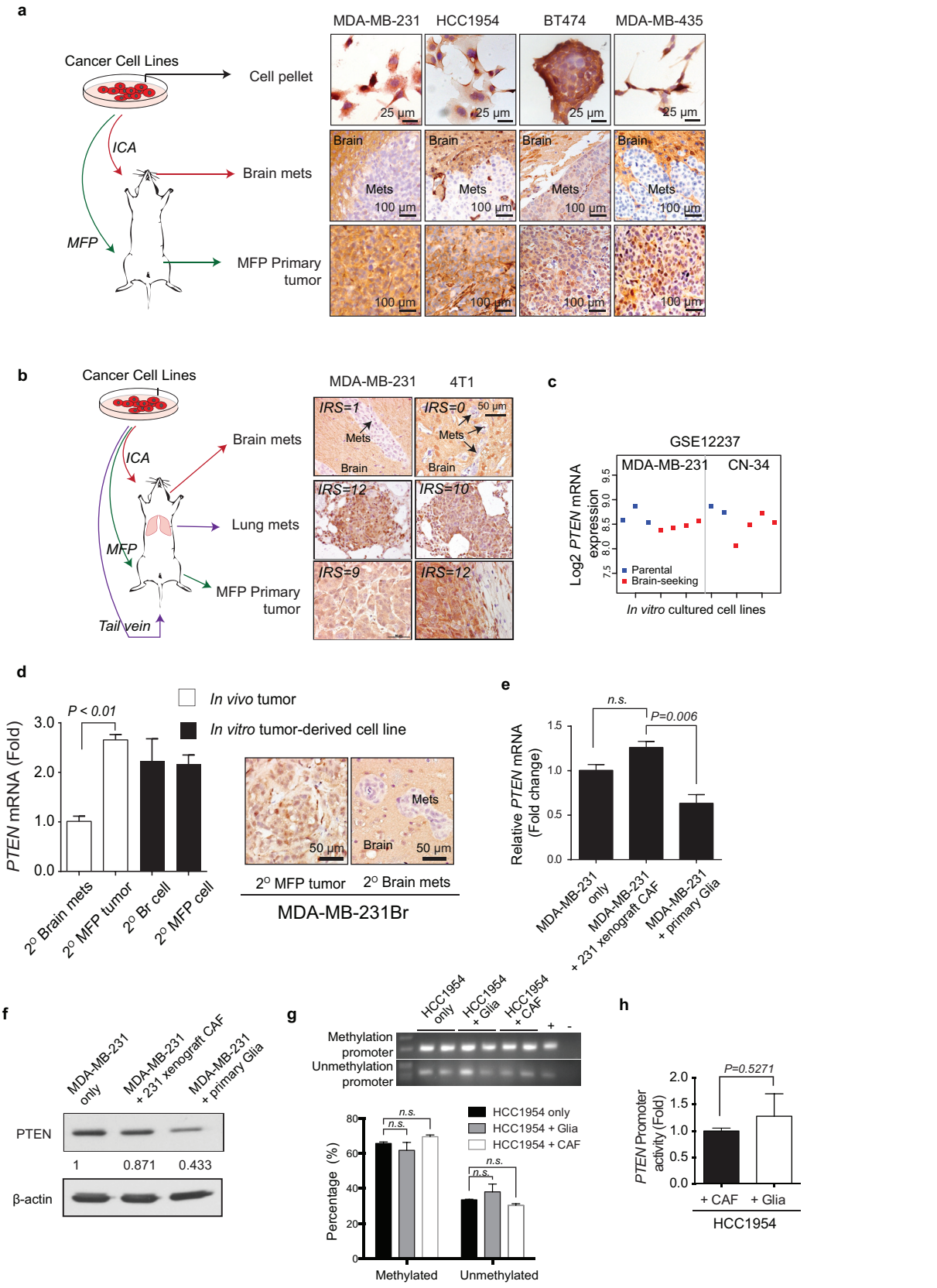
blocking with 1% goat serum. Slides underwent colour development with DAB and haematoxylin counterstaining. Ten visual fields from different areas of each tumour were evaluated by two pathologists independently (blinded to experiment groups). Positive IBA1 and Ki-67 staining in mouse tumours were calculated as the percentage of positive cells per field (%) and normalized by the total cancer cell number in each field. TUNEL staining was counted as the average number of positive cells per field (10 random fields). We excluded necrotic areas in the tumours from evaluation. Immunofluorescence was performed following the standard protocol recommended by Cell Signaling. In brief, after washing with PBS twice, cells were fixed with 4% formaldehyde. Samples were blocked with 5% normal goat serum in PBS for 1 h before incubation with a primary antibody cocktail overnight at 4 °C, washed, then incubated with secondary antibodies before examination using confocal microscope. Pathologists were blinded to the group allocation during the experiment and when assessing the outcome.

Bioinformatics and statistical analysis. Publicly available GEO data sets GSE14020, GSE19184, GSE2603, GSE2034 and GSE12276 were used for bioinformatics analysis. The top 2×10^4 verified probes were subjected to analysis. Differentially expressed genes between metastases from brain and other sites (primary or other metastatic organ sites) were analysed by SAM analysis in R statistical software. The 54 commonly downregulated genes in brain metastases from GSE14020 and GSE19184 were depicted as a heat-map by Java Treeview. For staining of patient samples, we calculated the correlation by Fisher's exact test. For survival analysis of GSE2603, the patient samples were mathematically separated into PTEN-low and -normal groups based on K-means ($K = 2$). Kaplan–Meier survival curves were generated by survival package in R. Multiple group IHC scores were compared by Chi-square test and Mantelhaen test in R. All quantitative experiments have been repeated using at least three independent biological repeats and are presented as mean ± s.e.m. or mean ± s.d.. Quantitative data were analysed either by one-way analysis of variance (ANOVA) (multiple groups) or t-test (two groups). $P < 0.05$ (two-sided) was considered statistically significant.

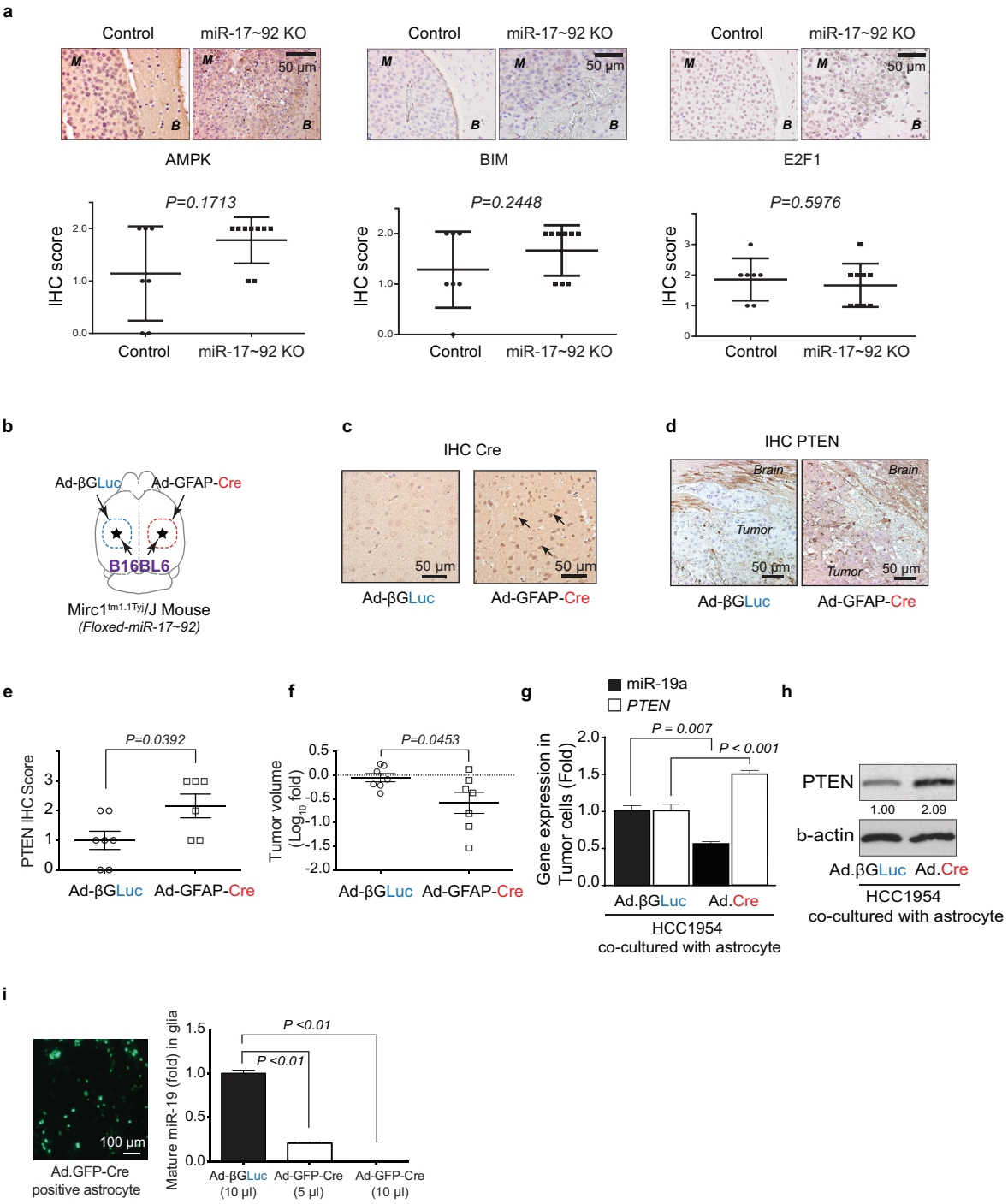
27. Lu, J. *et al.* 14-3-3ζ Cooperates with ErbB2 to promote ductal carcinoma in situ progression to invasive breast cancer by inducing epithelial-mesenchymal transition. *Cancer Cell* **16,** 195–207 (2009).
28. Nagata, Y. *et al.* PTEN activation contributes to tumor inhibition by trastuzumab, and loss of PTEN predicts trastuzumab resistance in patients. *Cancer Cell* **6,** 117–127 (2004).
29. Zhang, S. *et al.* Src family kinases as novel therapeutic targets to treat breast cancer brain metastases. *Cancer Res.* **73,** 5764–5774 (2013).
30. Gril, B. *et al.* Effect of lapatinib on the outgrowth of metastatic breast cancer cells to the brain. *J. Natl Cancer Inst.* **100,** 1092–1103 (2008).



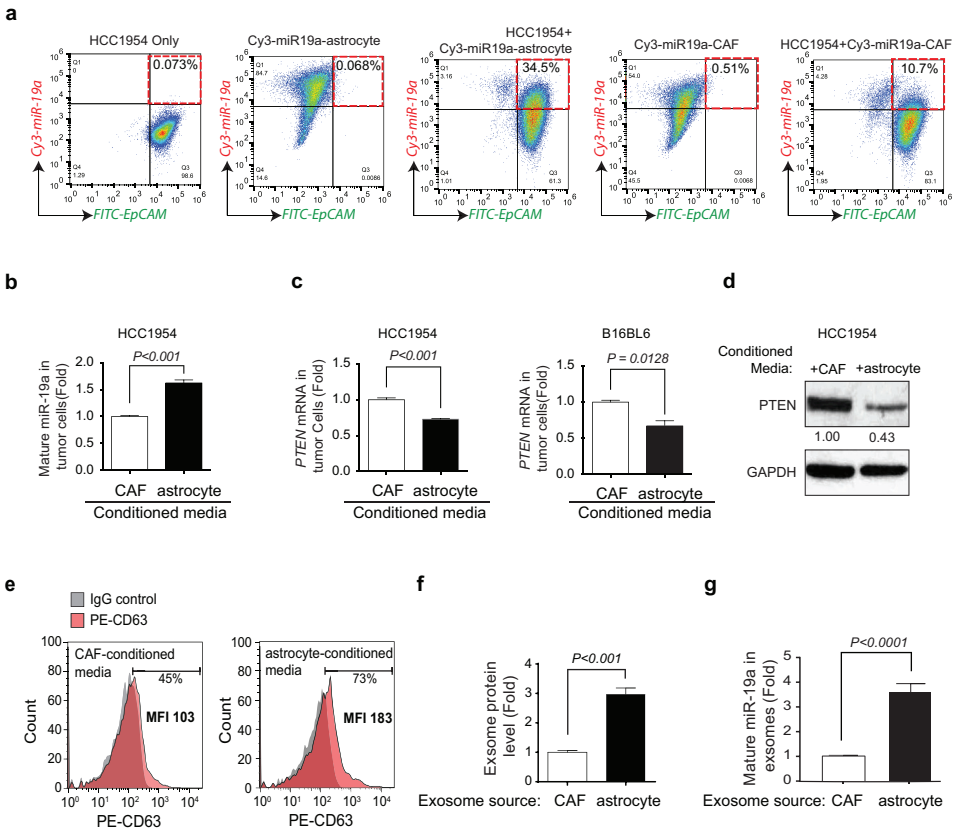
Extended Data Figure 1 | Organ-specific loss of *PTEN* in brain metastases. **a**, Schematics of microarray analyses. Patients' brain metastases exhibited a discrete gene expression profile with 650 genes significantly downregulated compared to bone or lung metastases (GSE14020). Cancer cells were injected into immunodeficient mice to produce orthotopic primary tumours (MDA-MB-231 cells for mammary tumour, PC14 for prostate tumour, A375SM for melanoma) and experimental brain metastases (all three lines). Brain metastases derived from these three cancer cell lines exhibited 2,161 commonly downregulated genes compared to their respective primary tumours (GSE19184). *PTEN* is one of only 54 commonly downregulated genes in brain metastases of both data sets. **b**, Heat-maps showing expression of 54 commonly downregulated genes (see **a**) in clinical brain metastases versus lung metastases and bone metastases. **c**, Heat-maps showing expression of the 54 genes (see **a**) in cell-line-induced primary tumours versus experimental brain metastases. **d**, Kaplan–Meier survival analyses showing no significant differences in brain metastasis-free survival between breast cancer patients with primary tumours expressing normal *PTEN* or low *PTEN* mRNA in GEO cDNA microarray set GSE2603 ($P = 0.74$). **e**, *PTEN* mRNA levels detected in primary breast tumours from patients with or without brain metastasis relapse. Three GEO cDNA microarray data sets (GSE2034, GSE2603 and GSE12276) with clinical annotation were analysed. Relative *PTEN* expression levels were compared by *t*-test.



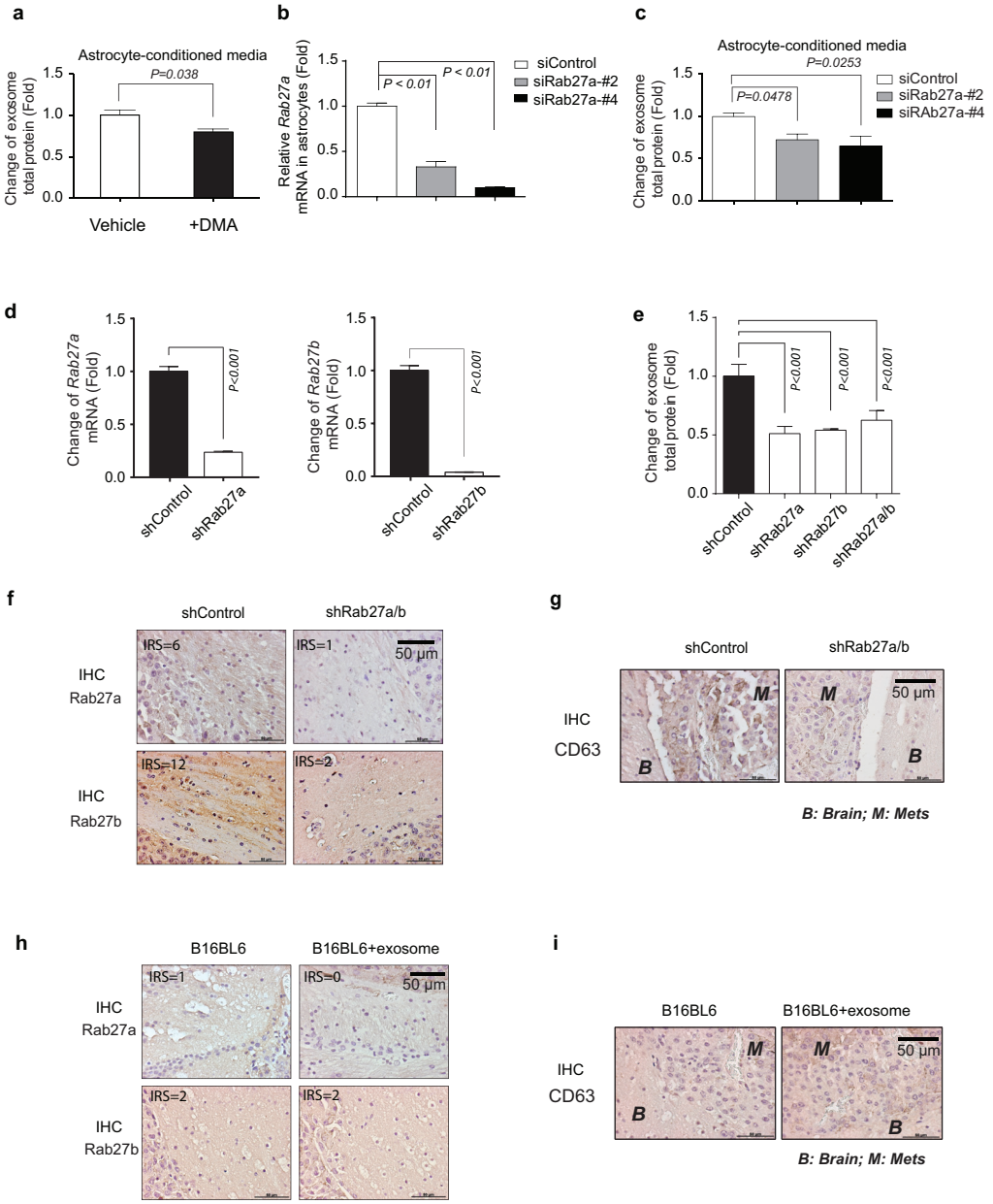
Extended Data Figure 2 | PTEN expression in different metastatic organ microenvironments and *in vitro* culture condition. **a**, Breast cancer cell lines (MDA-MB-231, HCC1954, BT474, and MDA-MB-435) were cultured and injected either to the MFP to form primary tumour or intracarotidly to form brain metastases. Cell pellets and tumour tissues were stained for PTEN expression using anti-PTEN antibodies as described previously[28]. **b**, IHC staining of PTEN in brain metastases, paired lung metastases and primary tumour derived from either MDA-MB-231 or 4T1 cells. PTEN expression level was analysed based on an IRS scoring system. **c**, *PTEN* mRNA levels between parental MDA-MB-231 and CN-34 breast cancer cell lines (blue) and their brain-seeking sublines (red). Normalized PTEN-specific probe intensity values were extracted from cDNA microarray data set GSE12237. Dot plot shows the mean probe intensity derived from independent RNA samples. **d**, *PTEN* qRT–PCR (mean ± s.e.m., *t*-test) and PTEN IHC in MDA-MB-231Br secondary tumours and cultured cells (3 biological replicates, with 3 technical replicates each). **e**, **f**, qRT–PCR (**e**) and western blot (**f**) analysis of *PTEN* mRNA expression (mean ± s.e.m., *t*-test) or protein expression in MDA-MB-231 cells after co-culture with either primary mouse CAFs isolated from MDA-MB-231 xenograft tumours or primary mouse glia isolated from mouse brain (3 biological replicates, with 3 technical replicates each). **g**, Representative methylation-specific PCR of *PTEN* promoter and quantification under co-culture with glia or CAF (mean ± s.e.m., *t*-test, 2 biological replicates, with 2 technical replicates each). **h**, *PTEN* promoter activity measured by luciferase reporter in HCC1954 cells after co-culture with either CAF or glia cells for 48 h (mean ± s.e.m., *t*-test, $P = 0.5271$, 3 biological replicates, with 3 technical replicates each).



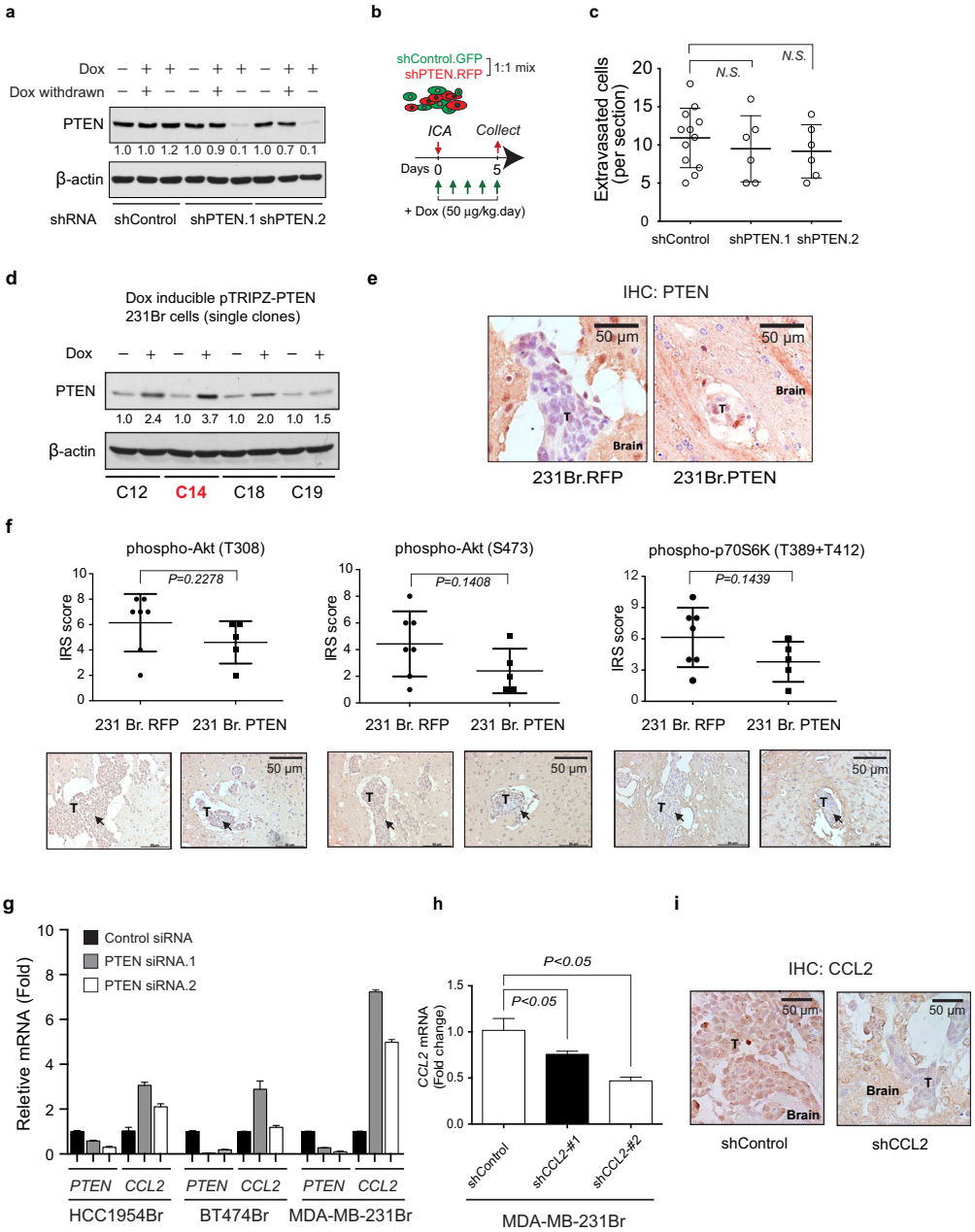
Extended Data Figure 3 | Cre-mediated depletion of PTEN-targeting microRNAs in astrocytes. **a**, IHC analyses of the expression of AMP-activated protein kinase (AMPK), pro-apoptotic protein BIM and transcription factor E2F1 (mean ± s.d., *t*-test) in brain metastasis tumours with/without pre-knocking out the miR-17~92 cluster in the brain microenvironment. B, brain tissue; M, brain metastases. **b**, Schematic of experimental design. The Ad-GFAP-Cre adenovirus was injected intracranially to the right hemisphere of the *Mirc1*$^{tm1.1Tyj}$*/J* mouse, and the control adenovirus (Ad-βGLuc) was injected intracranially to contralateral side of the brain. B16BL6 cells were then injected intracranially to both sides. **c**, IHC analysis of Cre expression in the brain astrocytes. **d**, IHC analysis of PTEN expression in the tumour cells. **e**, Quantification of PTEN expression in tumour cells (mean ± s.d., *t*-test). **f**, Quantification of intracranial tumour outgrowth by volume (mean ± s.e.m., *t*-test). **g**, qRT–PCR analyses of miR-19a and *PTEN* mRNA in tumour cell HCC1954 after 48 h co-culture with primary astrocytes from *Mirc1*$^{tm1.1Tyj}$*/J* mice pre-infected (48 h) with adenovirus (Ad-βGLuc or Ad-GFP-Cre) (mean ± s.e.m., *t*-test, 3 biological replicates, with 3 technical replicates each). **h**, Western blot of PTEN protein in the indicated tumour cells co-cultured as in **g**. **i**, Knockdown of miR-17~92 allele in cultured primary astrocytes. miR-17~92 cluster is flanked by *loxP* site in *Mirc1*$^{tm1.1Tyj}$*/J* mouse. Primary astrocytes were isolated from *Mirc1*$^{tm1.1Tyj}$*/J* mouse brain then infected by adenovirus encoding for βGLuc or GFP-Cre protein. Concentrated adenovirus particles of indicated volume (same MOI ~10^8 U ml^{-1}) encoding βGLuc or GFP-Cre proteins were added to 10^6 astrocytes. Left, representative image showing the infection efficiency. Right, bar diagram showing the relative miR-19a expression (one of the five miRNA genes in the miR-17~92 cluster) three days after adenovirus infection (mean ± s.e.m., *t*-test, 3 biological replicates, with 3 technical replicates each).



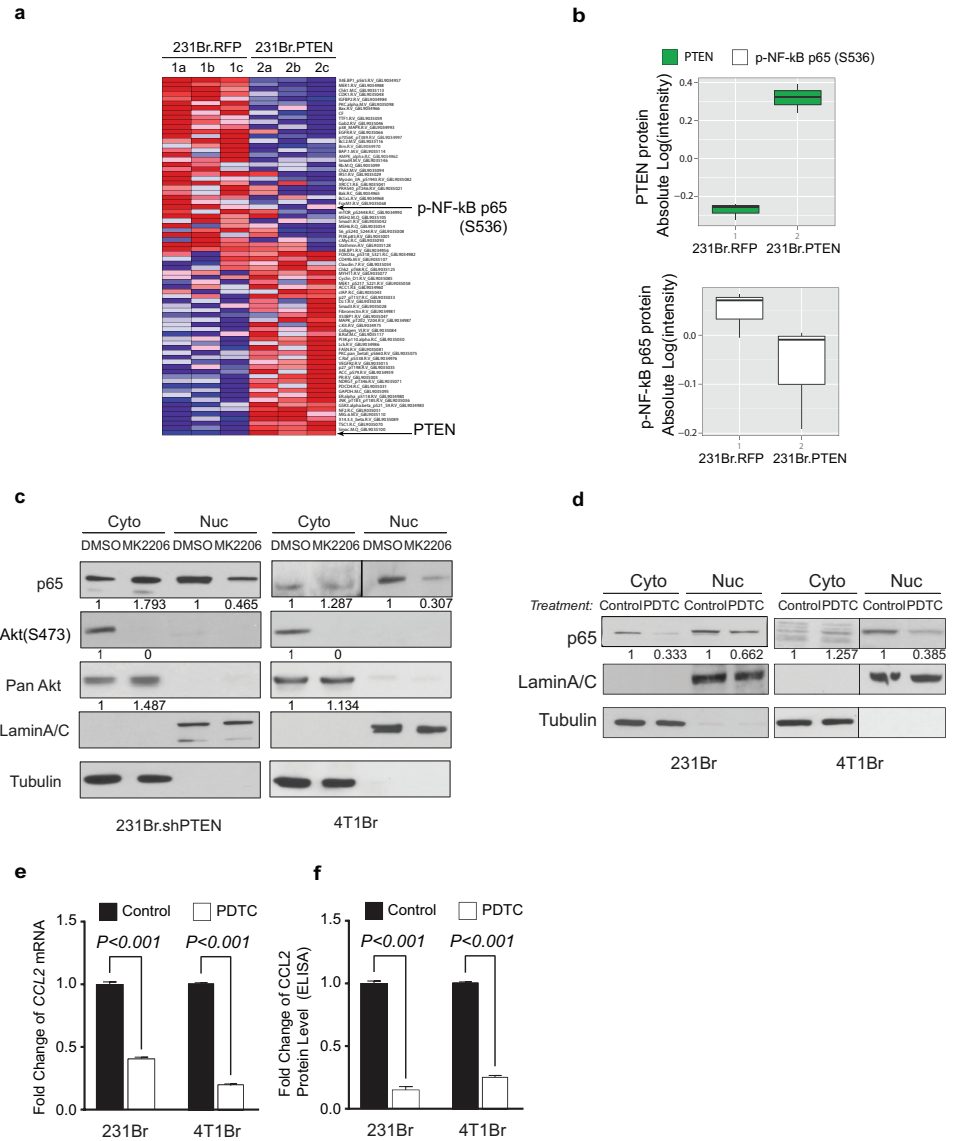
Extended Data Figure 4 | Contact-independent downregulation of PTEN in tumour cells by miR-19a from astrocyte-derived exosomes. **a**, Flow cytometric detection of Cy3-miR-19a and FITC-EpCAM in tumour cells 60 h after co-culture with Cy3-miR-19a-transfected astrocytes and CAFs. **b**, **c**, Tumour cells were co-cultured with conditioned media from astrocytes or CAFs for 60 h. RT–PCR analyses of the PTEN-targeting miR-19a level (**b**) and *PTEN* mRNA level (**c**) in tumour cells (mean ± s.e.m., *t*-test, 3 biological replicates, with 3 technical replicates each). **d**, Western blot detecting PTEN protein levels in HCC1954 cells after culture with conditioned media from either astrocytes or CAFs for 60 h. **e**, Flow cytometry detecting $CD63^+$ exosomes extracted from CAF- or astrocyte-conditioned media. **f**, Histogram showing the exosome protein level detected from CAF- and astrocyte-conditioned media normalized by cell number (mean ± s.e.m., *t*-test, $P < 0.0001$, 3 biological replicates, with 3 technical replicates each). **g**, RT–PCR analyses of miR-19a level in exosomes extracted from CAF- or astrocyte-conditioned media normalized by equal cell numbers (mean ± s.e.m., *t*-test, $P < 0.0001$, 3 biological replicates, with 3 technical replicates each).



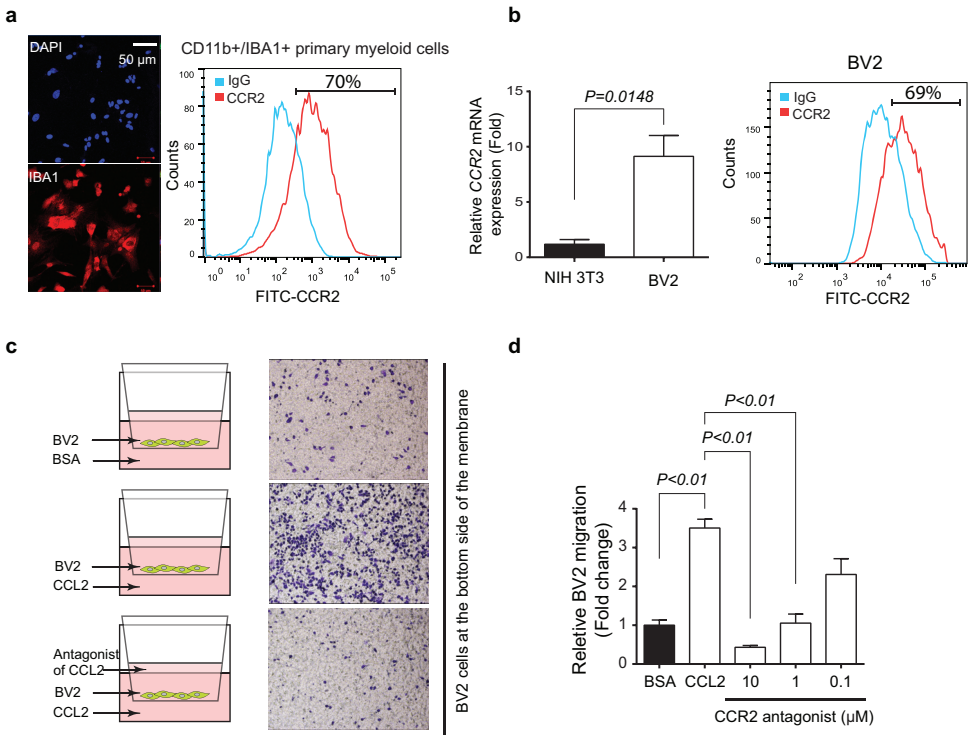
Extended Data Figure 5 | Inhibition of exosome release by DMA, *Rab27a* siRNA or *Rab27* shRNAs. **a**, Exosome-releasing inhibitor (DMA) treatment reduced exosome secretion from astrocytes compared to vehicle treated astrocytes. Astrocytes were treated with DMA (25 μg ml^{-1}) or vehicle for 4 h; exosomes were concentrated from astrocyte-conditioned media and total proteins from exosomes were examined by BCA assay (normalized to total cell numbers) (mean ± s.e.m., *t*-test, $P = 0.038$, 3 biological replicates, with 3 technical replicates each). **b**, Knockdown of *Rab27a* in astrocytes by siRNA. Two siRNAs targeting mouse *Rab27a* were transiently transfected into astrocytes, and the *Rab27a* mRNA level was examined by RT–PCR 48 h after transfection (mean ± s.e.m., *t*-test, $P < 0.01$, 3 biological replicates, with 3 technical replicates each). **c**, Knocking down *Rab27a* in astrocytes inhibited exosome release. Forty-eight hours after *Rab27a*-targeting siRNAs were transfected, exosomes were collected from astrocyte-conditioned media and total proteins from exosomes were examined by BCA assay (normalized to total cell numbers) (mean ± s.e.m., *t*-test, 3 biological replicates, with 3 technical replicates each). **d**, Histogram showing relevant changes of *Rab27a* and *Rab27b* mRNA level in primary astrocytes infected with pLKO.shRab27a or pLKO.shRab27b virus (mean ± s.e.m., *t*-test, $P < 0.001$, 3 biological replicates, with 3 technical replicates each). **e**, Change of exosome protein level detected in the conditioned media from astrocytes infected by pLKO.shRab27a or pLKO.shRab27b virus by BCA assay (normalized to total cell numbers) (mean ± s.e.m., *t*-test, $P < 0.001$, 3 biological replicates, with 3 technical replicates each). **f**, **g**, IHC analysis showing the expression level of Rab27a and Rab27b (**f**) and exosome marker expression CD63 (**g**) in the brain tissue derived from mice injected with control lentivirus or *Rab27a/b* shRNA lentiviruses and subsequently intracranially injected with B16BL6 cells. **h**, **i**, IHC analysis showing the expression level of Rab27a and Rab27b (**h**) and exosome marker expression CD63 (**i**) in the brain tissue derived from mice injected with *Rab27a/b* shRNA lentiviruses and subsequently intracranially injected with B16BL6 cells and vehicle at the left side or B16BL6 cells and astrocyte-derived exosomes at the right side.



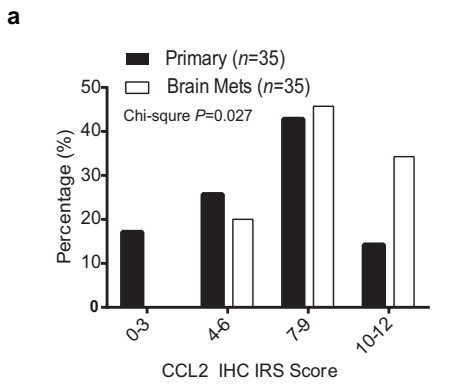
Extended Data Figure 6 | Brain extravasation of MDA-MB-231 parental cells with or without induction of doxycycline-inducible *PTEN* shRNA knockdown, PTEN expression and *CCL2* shRNA knockdown. **a**, Western blot showing PTEN expression levels after treating MDA-MD-231 cells with doxycycline. MDA-MD-231 cells were stably infected with inducible shRNA expression vectors (pTRIPZ-control-shGFP as control and pTRIPZ-shRNA-RFP for *PTEN* shRNA). Doxycycline (1 μg ml^{-1}) was added to induce shRNA expression for 5 days. As indicated, doxycycline was withdrawn in some samples for another 5 days before analysis. **b**, Schematics of *in vivo* extravasation assay. shControl-GFP and shPTEN-RFP cells were mixed at a 1:1 ratio. In total, 200,000 cells were ICA injected into mice, and doxycycline (50 μg kg^{-1}) was given intraperitoneally daily. Brains were collected 5 days after ICA injection. **c**, Dot plot of extravasated cell counts 5 days after ICA injection of indicated MDA-MB-231 sublines. Tumour-bearing brains were collected and sectioned into 100 μm coronal slices. Extravasated tumour cells were counted under the fluorescence microscope (mean ± s.d., *t*-test). **d**, MDA-MB-231Br single cells were expanded into subclones (C12, C14, C18 and C19), which were transfected with doxycycline-inducible pTRIPZ-RFP or pTRIPZ-PTEN. 48 h after doxycycline (1 μg ml^{-1}) treatment, PTEN induction was tested by western blotting. The C14 clone was used for further *in vivo* assays (see **e**, **f** and Fig. 4a). **e**, IHC staining of induced PTEN expression in brain metastases derived from mice injected with MDA-MB-231Br (231Br-RFP or 231Br-PTEN) cells. **f**, IHC analysis of PTEN downstream signalling pathway, including phosphorylated pAkt(T308), pAkt(S473) and pP70S6K(T389+T412) in brain metastases from mice injected with 231Br-RFP or 231Br-PTEN cells. Top, dot plot of IHC data quantification by IRS (mean ± s.d., *t*-test); bottom, representative IHC staining data. **g**, Histograms of *PTEN* and *CCL2* mRNA levels (mean ± s.e.m., *t*-test) in indicated cancer cell lines 48 h after transfection with control or *PTEN* siRNAs (3 biological replicates, with 3 technical replicates each). **h**, Histogram showing the inducible CCL2 knockdown. MDA-MB-231Br cells were stably infected with pTRIPZ-inducible *CCL2* shRNAs. 48 h after doxycycline (1 μg ml^{-1}) treatment, *CCL2* mRNA was examined by RT–PCR (mean ± s.e.m., *t*-test, 3 biological replicates, with 3 technical replicates each). **i**, Doxycycline-induced CCL2 knockdown in brain metastases. Mice were ICA injected with MDA-MB-231Br cells containing control or *CCL2* shRNAs. Doxycycline (50 μg kg^{-1}) was given to mice intraperitoneally daily after injection. IHC staining of CCL2 expression levels in brain metastases derived from MDA-MB-231Br cells. T, brain metastasis tumours at day 30 after ICA injection.



Extended Data Figure 7 | PTEN-regulated CCL2 expression through the NF-κB pathway. **a**, Heat-map showing differentially expressed protein markers of reverse-phase protein array analysis. MDA-MB-231Br cells were stably infected with pTRIPZ-RFP or pTRIPZ-PTEN (231Br-RFP or 231Br-PTEN) and induced by doxycycline (1 μg ml^{-1}) for 48 h. **b**, Box chart showing the absolute intensity of PTEN and NF-κB p65(S536). **c**, Western blot analysis of NF-κB p65 nuclear translocation, after cells were treated with Akt inhibitor MK2206 (10 μg ml^{-1}) 24 h before separation into cytosolic (Cyto) and nuclear (Nuc) fractions. **d**, Western blot analysis of NF-κB p65 nuclear translocation, after cells were treated with NF-κB inhibitor PDTC (0.2 mM) 16 h before separation into cytosolic and nuclear fractions. **e**, Relative *CCL2* mRNA expression after NF-κB inhibitor PDTC treatment analysed by qRT–PCR (mean ± s.e.m., *t*-test, 3 biological replicates, with 3 technical replicates each). Cells were treated with PDTC (0.2 mM) for 16 h. **f**, Relative CCL2 protein expression after PDTC treatment analysed by ELISA (mean ± s.e.m., *t*-test, 3 biological replicates, with 3 technical replicates each). Cells were treated as in **e**.



Extended Data Figure 8 | CCR2-mediated IBA1$^+$ myeloid cell directional migration. **a**, Co-expression of IBA1 and CCR2 on myeloid cells freshly isolated from mouse brain by CD11b beads. Representative immunofluorescence staining of IBA1 (left). FACS analysis of CD11b$^+$ cells for CCR2 expression. **b**, Relative CCR2 expression in the BV2 microglia cell line compared with NIH3T3 fibroblasts. *CCR2* mRNA level analysed by qRT–PCR (mean ± s.e.m., *t*-test, 3 biological replicates, with 3 technical replicates each) (left) and protein expression analysed by FACS (right) **c**, Transwell migration assay examining the directional migration of BV2 cells towards CCL2. In total, 10^5 BV2 cells were seeded in the top chamber of the transwell units, and CCL2 or BSA (20 ng ml^{-1}) was added into serum-free media in the bottom chamber. The migrated cell numbers were counted at 24 h. Next, CCR2 antagonists with different concentrations (10 μM, 1 μM and 0.1 μM) were added into the top chamber with BV2 cells, and CCL2 (20 ng ml^{-1}) was added into serum-free media in the bottom chamber. The migrated cell numbers were counted at 24 h. **d**, Quantification of BV2 cell migration assay (mean ± s.e.m., *t*-test, 3 biological replicates, with 3 technical replicates each).



b

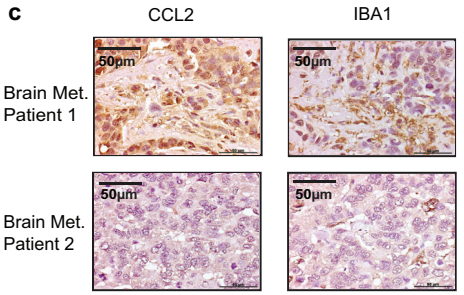
Primary Breast Tumor

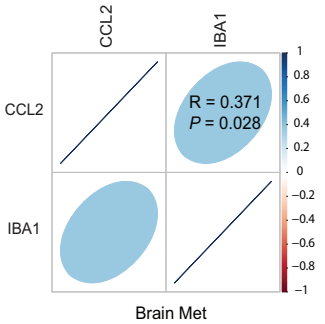
CCL2 \ PTEN	0-3 (%)	4-12 (%)	Total	*P*-value *
0-6 (%)	5 (14.3%)	10 (28.6%)	15	
7-12 (%)	8 (22.8%)	12 (34.3%)	20	=0.686
Total	13	22	35	

Brain Met

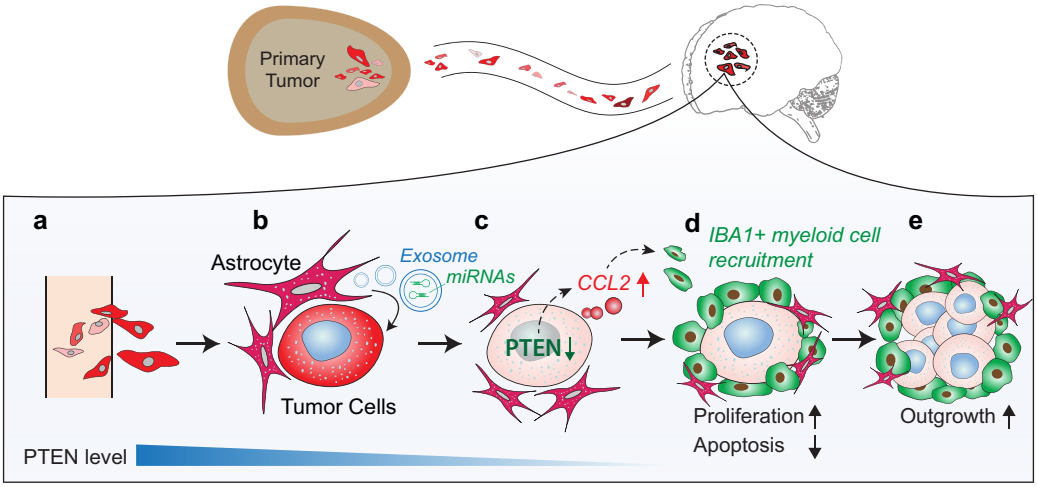
CCL2 \ PTEN	0-3 (%)	4-12 (%)	Total	*P*-value *
0-6 (%)	7 (20.0%)	0 (0%)	7	
7-12 (%)	18 (51.4%)	10 (28.6%)	28	=0.061
Total	25	10	35	

* Chi-square test





Extended Data Figure 9 | The association between PTEN, CCL2 expression and recruitment of IBA1$^+$ myeloid cells in patients' brain metastases and matched primary breast tumours. **a**, Summary histogram of CCL2 protein levels in primary breast tumours and matched brain metastases from 35 patients. Chi-square test was used to compare the IHC score in primary breast tumours versus matched brain metastases. $P < 0.05$ is defined as significantly different. **b**, Tables showing IHC scores of PTEN and CCL2 expression in primary breast tumours and matched brain metastases. **c**, Representative IHC staining of CCL2 proteins and IBA1$^+$ myeloid cells in patients' brain metastases, and the correlation plot showing the Pearson correlation between CCL2 and IBA1 staining in patients' brain metastases ($R = 0.371$, $P = 0.028$).



Extended Data Figure 10 | PTEN loss induced by astrocyte-derived exosomal microRNA primes brain metastasis outgrowth via functional cross-talk between disseminated tumour cells and brain metastatic microenvironment. Top, disseminated tumour cells extravasate into the brain. **a**–**c**, Exosomes secreted by astrocytes in the brain microenvironment transfer PTEN-targeting miRNA into extravasated brain metastatic tumour cells, leading to PTEN downregulation in tumour cells. **c**, **d**, PTEN loss in brain metastatic tumour cells increases their CCL2 secretion, facilitating the recruitment of $IBA1^{+}/CCR2^{+}$ myeloid cells at the micrometastasis site. **d**, **e**, The recruited $IBA1^{+}$ myeloid cells enhance proliferation and inhibit apoptosis of metastatic tumour cells, and promote metastatic outgrowth.

LETTER

doi:10.1038/nature15548

Autophagy mediates degradation of nuclear lamina

Zhixun Dou[1], Caiyue Xu[1], Greg Donahue[1], Takeshi Shimi[2], Ji-An Pan[3], Jiajun Zhu[1], Andrejs Ivanov[4]†, Brian C. Capell[1], Adam M. Drake[1], Parisha P. Shah[1], Joseph M. Catanzaro[3], M. Daniel Ricketts[5], Trond Lamark[6], Stephen A. Adam[2], Ronen Marmorstein[5,7,8], Wei-Xing Zong[3], Terje Johansen[6], Robert D. Goldman[2], Peter D. Adams[4] & Shelley L. Berger[1]

Macroautophagy (hereafter referred to as autophagy) is a catabolic membrane trafficking process that degrades a variety of cellular constituents and is associated with human diseases[1–3]. Although extensive studies have focused on autophagic turnover of cytoplasmic materials, little is known about the role of autophagy in degrading nuclear components. Here we report that the autophagy machinery mediates degradation of nuclear lamina components in mammals. The autophagy protein LC3/Atg8, which is involved in autophagy membrane trafficking and substrate delivery[4–6], is present in the nucleus and directly interacts with the nuclear lamina protein lamin B1, and binds to lamin-associated domains on chromatin. This LC3–lamin B1 interaction does not downregulate lamin B1 during starvation, but mediates its degradation upon oncogenic insults, such as by activated RAS. Lamin B1 degradation is achieved by nucleus-to-cytoplasm transport that delivers lamin B1 to the lysosome. Inhibiting autophagy or the LC3–lamin B1 interaction prevents activated RAS-induced lamin B1 loss and attenuates oncogene-induced senescence in primary human cells. Our study suggests that this new function of autophagy acts as a guarding mechanism protecting cells from tumorigenesis.

Several mammalian autophagy proteins are present in the nucleus, including LC3 (refs 7, 8), Atg5 (ref. 9), and Atg7 (ref. 10). However, whether nuclear LC3 is involved in degrading nuclear components is not understood. We investigated LC3 distribution by subcellular fractionation of primary human IMR90 cells and found a substantial amount of endogenous LC3 and a small amount of lipidated LC3-II in the nucleus (Fig. 1a). We used bacterially purified glutathione *S*-transferase (GST)–LC3B (hereafter 'LC3', unless specified otherwise) to pull down the nuclear fraction (Fig. 1b). One protein that we found to interact with LC3 is the nuclear lamina protein lamin B1 (Fig. 1b). The nuclear lamina is a fibrillar network located beneath the nuclear envelope whose major components are the four nuclear lamin isoforms: lamins B1, B2, and A/C, and their associated proteins[11]. Nuclear lamina provides the nucleus with mechanical strength and regulates higher-order chromatin organization, modulating gene expression and silencing[11]. In contrast to lamin B1, lamins A/C and lamin B2 bind poorly, if at all, to LC3 (Fig. 1b). We detected a direct interaction of purified lamin B1 (Extended Data Fig. 1a) with LC3B (Fig. 1c) and other members of the Atg8 protein family, including LC3A, LC3C, and GABARAP (Extended Data Fig. 1b, c). Co-immunoprecipitation (co-IP) revealed that LC3–lamin B1 interaction occurs at the endogenous level in the nucleus (Fig. 1d, e and Extended Data Fig. 1d). Lipidated LC3-II is involved in mediating lamin B1 interaction (Fig. 1d and Extended Data Fig. 1e–g), and the LC3 G120A lipidation-deficient mutant showed impaired binding to lamin B1 (Fig. 1f). A bimolecular fluorescence complementation (BiFC) assay[12] showed that LC3–lamin B1 interaction happens at the nuclear lamina and is dependent on LC3 lipidation (Extended Data Fig. 1h–j). Together, these data suggest that LC3 directly interacts with lamin B1, and that LC3 lipidation facilitates this interaction, possibly by tethering LC3 to the inner nuclear membrane where the interaction with nuclear lamina occurs.

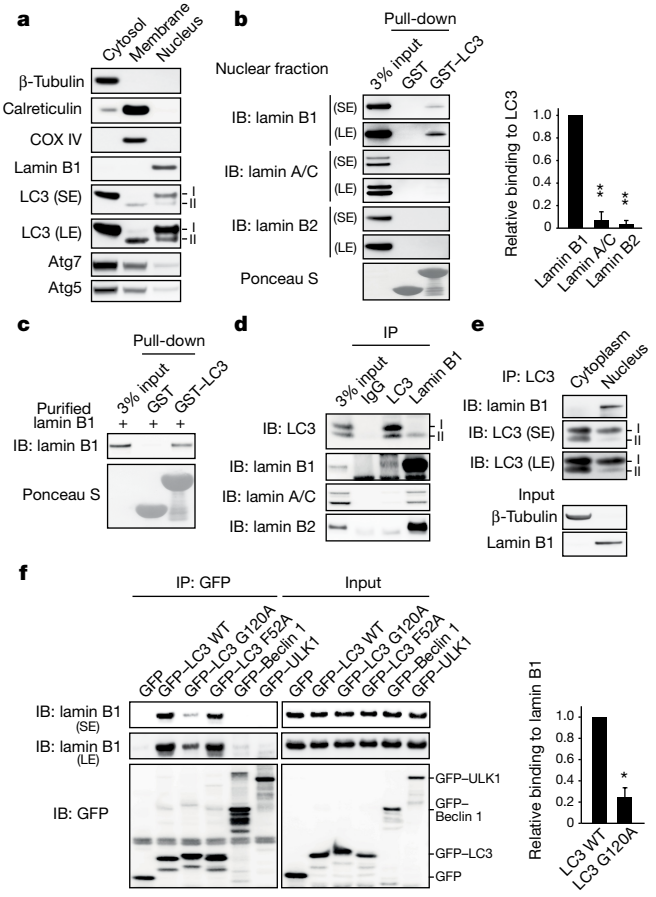


Figure 1 | LC3 interacts with nuclear lamina protein lamin B1. **a**, Proliferating young IMR90 cells were subjected to subcellular fractionation and immunoblotting. SE, short exposure; LE, long exposure. **b**, The nuclear fraction of IMR90 cells was pulled down with bacterially purified GST or GST–LC3. **c**, GST–LC3 pull-down of purified lamin B1 protein. **d**, Endogenous immunoprecipitation in IMR90 cells. **e**, LC3 immunoprecipitation of IMR90 fractions. **f**, HEK293T cells were transfected and subjected to GFP immunoprecipitation and immunoblotting. Bars, mean ± s.e.m.; $n = 3$; $*P < 0.001$. $**P < 0.0001$; one-way analysis of variance (ANOVA) coupled with Tukey's *post hoc* test (**b**); unpaired two-tailed Student's *t*-test (**f**). Uncropped blots are in Supplementary Figure 1.

[1]Epigenetics Program, Department of Cell and Developmental Biology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA. [2]Department of Cell and Molecular Biology, Feinberg School of Medicine, Northwestern University, Chicago, Illinois 60611, USA. [3]Department of Molecular Genetics and Microbiology, Stony Brook University, Stony Brook, New York 11794, USA. [4]Institute of Cancer Sciences, University of Glasgow and Beatson Institute for Cancer Research, Glasgow G61 1BD, UK. [5]Department of Biochemistry & Biophysics, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA. [6]Molecular Cancer Research Group, Institute of Medical Biology, University of Tromsø – The Arctic University of Norway, 9037 Tromsø, Norway. [7]Department of Chemistry, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA. [8]Abramson Family Cancer Research Institute, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA. †Present address: Centre for Haemato-Oncology, Barts Cancer Institute, Queen Mary University of London, Charterhouse Square, London EC1M 6BQ, UK.

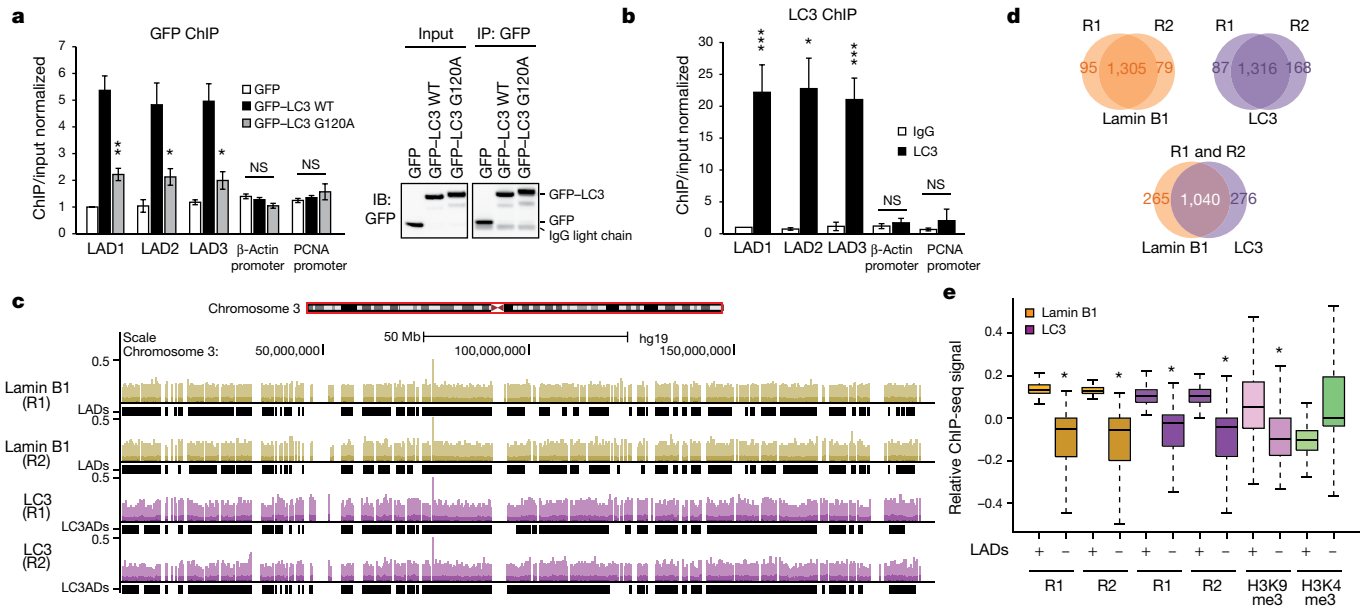


Figure 2 | LC3 associates with LADs on chromatin. a, IMR90 cells stably expressing GFP-tagged constructs were subjected to GFP ChIP–quantitative polymerase chain reaction (qPCR). Uncropped blots are in Supplementary Figure 1. **b**, LC3 ChIP–qPCR. Bars, mean ± s.e.m.; $n = 3$; $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.005$; NS, non-significant; unpaired two-tailed Student's *t*-test. **c–e**, ChIP-sequencing analyses in proliferating IMR90 cells. **c**, Representative tracks over the whole of chromosome 3, for both replicates. **d**, Overlap of LADs and LC3ADs between two replicates and lamin B1 and LC3. **e**, ChIP-seq enrichment over LADs (+) and randomly selected non-LADs control regions (−). One-sided Wilcoxon test; $^{*}P < 2.2 \times 10^{-16}$, $P = 1$ for H3K4me3.

Lamin B1 associates with transcriptionally inactive heterochromatin domains called LADs (lamin-associated domains)[11,13]. We used chromatin immunoprecipitation (ChIP) to investigate the association of LC3 with LADs. ChIP of LC3 showed that in its lipidated form, LC3 associates with LADs but poorly with euchromatin regions, such as β-actin and PCNA promoters, similarly to that of lamin B1 (Fig. 2a, b and Extended Data Fig. 2a–c). We then performed endogenous lamin B1 and LC3 ChIP followed by genome-wide sequencing (ChIP-seq), done in two independent biological replicates, R1 and R2 (Fig. 2c for whole chromosome 3 and a zoom-in window in Extended Data Fig. 2d). We used enriched domain detector (EDD), an algorithm that detects wide enrichment domains[14] to define LADs and LC3-associated domains (LC3ADs) (Fig. 2c and Extended Data Fig. 2d, black rectangles beneath the tracks). Analyses of lamin B1 and LC3 ChIP-seq revealed high reproducibility between R1 and R2 over LADs and LC3ADs (Fig. 2d, top two panels, and Extended Data Fig. 2e, f); LADs defined here correlate well with previously identified LADs from lamin B1 ChIP-seq[15,16] and DamID[13] (Extended Data Fig. 2g). We further found that LADs and LC3ADs significantly overlap (Fig. 2d, bottom panel; permutation test $P < 0.001$, 1,000 iterations). Comparing LADs with an equal number of size-matched and randomly selected non-LAD control regions, we observed that both lamin B1 and LC3 are strongly enriched in LADs, for both replicates (Fig. 2e; permutation test for LC3: $P < 0.01$, 100 iterations, for both replicates). A similar enrichment is also detected over LC3ADs (Extended Data Fig. 2h). As expected, Lys9 trimethylation on histone H3 (H3K9me3) is highly enriched in LADs (Fig. 2e, permutation test $P < 0.01$, 100 iterations), whereas H3K4me3 is relatively depleted (Fig. 2e, permutation test $P = 1$, 100 iterations). We also found that both lamin B1 and LC3 from our ChIP-seq are strongly enriched in LADs mapped by other published studies[13,15] relative to non-LAD control regions (Extended Data Fig. 2i), in line with our findings from Fig. 2e. Collectively, these results indicate that LC3 associates with LADs on chromatin at the genome-wide scale.

Next, we examined the biological functions of this interaction, and found that neither starvation nor rapamycin treatment downregulates lamin B1 protein (Fig. 3a), suggesting that autophagy does not degrade lamin B1 during starvation. One scenario that involves lamin B1 loss is oncogenic insult, such as induced by oncogenic *RAS*[17–19]. In fact, most primary cells and tissues cope with oncogenic *RAS* activity by inducing cellular senescence, a stable cell-cycle arrest that serves as a potent tumour suppressive mechanism[20,21]. We and others have shown that lamin B1, but not lamins A/C or B2, is dramatically downregulated during oncogene-induced senescence[17–19]. Importantly, autophagy is

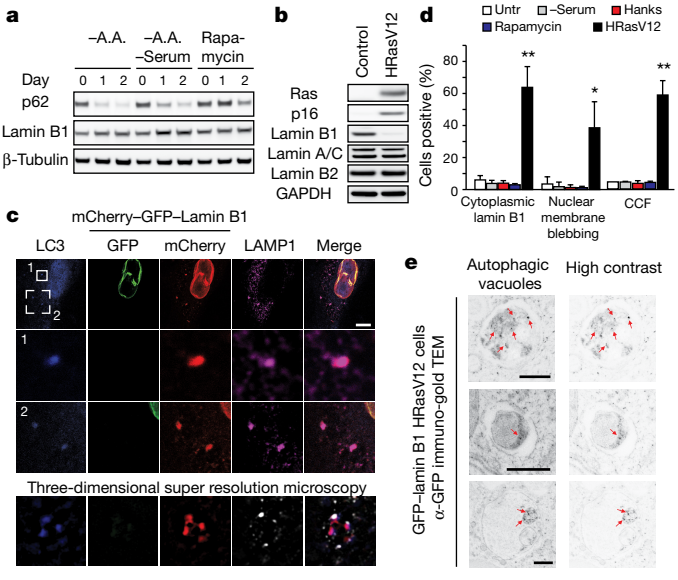


Figure 3 | Lamin B1 is an autophagy substrate in response to oncogene activation. a, **b**, Primary IMR90 cells were treated as indicated and subjected to immunoblotting. AA, amino acids. Uncropped blots are in Supplementary Figure 1. **c**, IMR90 cells stably expressing mCherry–GFP–lamin B1 and HRasV12 were stained with LC3 and LAMP1 antibodies, and analysed by confocal or three-dimensional super-resolution microscopy. Scale bar, 10 μm. **d**, mCherry–GFP–lamin B1 IMR90 cells were treated as indicated. Bars, mean ± s.d.; $n = 4$; $^{*}P < 0.01$, $^{**}P < 0.001$; one-way ANOVA coupled with Tukey's *post hoc* test. **e**, Immuno-TEM analysis of IMR90 cells stably expressing GFP–lamin B1 and HRasV12. Gold nanoparticles are indicated by arrows and highlighted on the right. Scale bar, 500 nm.

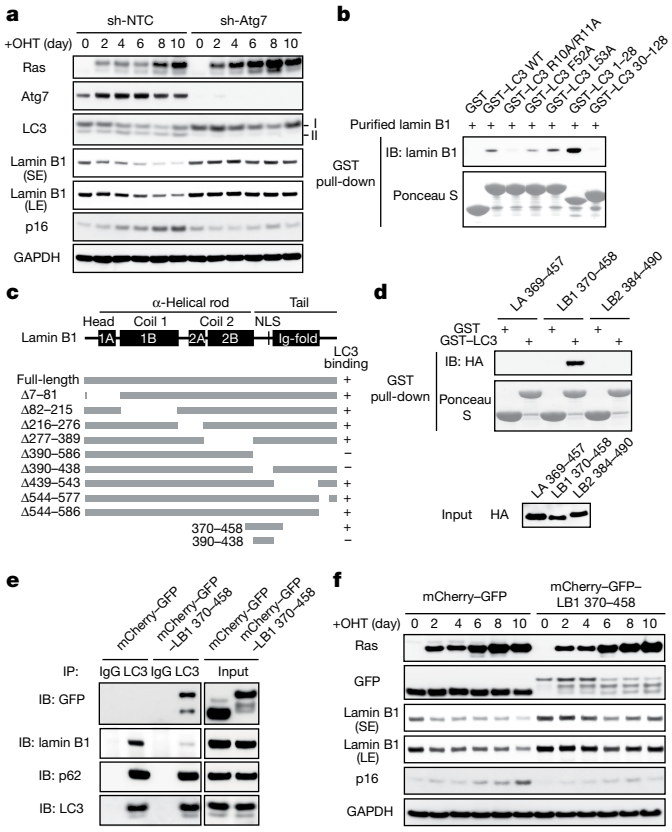


Figure 4 | Inhibiting autophagy or the LC3–lamin B1 interaction impairs lamin B1 degradation. a, ER:HRasV12 IMR90 cells stably expressing non-targeting control (sh-NTC) or sh-Atg7 hairpin were induced by OHT (4-hydroxytamoxifen) and analysed by immunoblotting. **b**, Purified lamin B1 protein was subjected to pull-down of GST–LC3B wild type or mutants. **c**, Schematic illustration of lamin B1 mutants in binding to LC3. **d**, Regions from lamin A, B1, and B2 were subjected to GST–LC3B pull-down. **e**, HEK293T transfected were subjected to LC3 immunoprecipitation. **f**, ER:HRasV12 IMR90 cells were induced by OHT and analysed by immunoblotting. Uncropped blots are in Supplementary Figure 1.

upregulated during oncogene-induced senescence, and is required for the mitosis-to-senescence transition[22,23]. We thus hypothesized that activated oncogenes trigger autophagic degradation of lamin B1 in primary human cells.

Consistent with previous findings[17,18], primary, but not immortalized, human cells show downregulation of lamin B1 but not other lamin isoforms (Fig. 3b and Extended Data Fig. 3a). Although starvation does not alter lamin B1 nuclear lamina localization, HRasV12 expression induces nuclear membrane blebbing and cytoplasmic lamin B1 signals (Extended Data Fig. 3b). Transmission electron microscopy (TEM) analysis of HRasV12-expressing cells confirmed the induction of autophagosomes, reduction of perinuclear heterochromatin, and induction of nuclear membrane blebs (Extended Data Fig. 3c–e). Unlike yeast piecemeal microautophagy, in which nuclear blebs directly contact cytoplasmic autophagic vacuoles[24], the nuclear blebs in human senescent cells are morphologically distinct and do not directly contact these vacuoles (Extended Data Fig. 3c–e).

We further used an mCherry–GFP–lamin B1 construct to investigate the hypothesis that lamin B1 is degraded by the autophagy–lysosome pathway. Here, a yellow signal (due to merged mCherry and GFP) indicates that the fusion protein is in a neutral pH environment, whereas a red signal (due to quenching of GFP) indicates that the protein has entered acidic lysosomes[25,26]. mCherry–GFP–lamin B1 showed a merged yellow nuclear peripheral pattern in control cells, but displayed cytoplasmic red-only bodies in HRasV12-expressing cells (Extended Data Fig. 3f). Inhibiting lysosomal acidification by bafilomycin A1 prevents GFP quenching and results in merged yellow signals in the cytoplasm (Extended Data Fig. 3g). Furthermore, we co-stained with antibodies against LC3 and LAMP1, and found that the cytoplasmic mCherry-only lamin B1 bodies stain positively for endogenous LC3 and LAMP1 (Fig. 3c). Super-resolution microscopy analysis revealed that the cytoplasmic lamin B1 and LC3 co-localizes within the LAMP1-decorated vesicle (Fig. 3c and Extended Data Fig. 4a). Cytoplasmic lamin B1 and nuclear membrane blebs are specifically induced by HRasV12, but not by starvation or rapamycin treatment (Fig. 3d). In addition, we performed live-cell imaging on mCherry–GFP–lamin B1-expressing HRasV12 IMR90 cells, and confirmed a nucleus-to-cytoplasm transport process, through nuclear membrane blebbing, which then leads to lamin B1 degradation in the cytoplasm (Extended Data Fig. 4b).

Cytoplasmic lamin B1 in HRasV12 cells is reminiscent of the cytoplasmic chromatin fragments (CCF) that we previously described in senescent cells, which are fragments of heterochromatin budded off from the nuclei[19]. Consistent with the behaviour of lamin B1, we found cytoplasmic DAPI (4′,6-diamidino-2-phenylindole) specifically appearing in response to HRasV12 (Fig. 3d). The cytoplasmic DAPI staining bodies are positive for H3K27me3 and H3K9me3, and co-localize with LC3 and lamin B1 (Extended Data Fig. 5a–c). Immuno-TEM analysis revealed that lamin B1 specifically localizes at the nuclear lamina in control cells (Extended Data Fig. 5d, left), whereas HRasV12-expressing cells showed decreased presence of lamin B1 at the nuclear lamina, and the appearance inside autophagosomes and autolysosomes (Fig. 3e and Extended Data Fig. 5d, right). Taken together, these data indicate that lamin B1 is an autophagy substrate upon oncogenic insult, which, through a nucleus-to-cytoplasm transport process, leads to its autophagic degradation in the cytoplasm.

We subsequently investigated the consequence of autophagy inhibition. Knockdown of Atg7 impairs the downregulation of lamin B1 protein in HRasV12 cells (Fig. 4a and Extended Data Fig. 6a). Lamin B1 messenger RNA (mRNA) has been shown to decrease upon HRasV12 expression[17,18]. Here the mRNA of lamin B1 is reduced both in control and in Atg7 knockdown cells (Extended Data Fig. 6b), whereas the protein level of lamin B1 is maintained in Atg7-deficient cells (Fig. 4a). These data suggest that lamin B1 is downregulated both at mRNA and at protein levels, and are consistent with the observation that nuclear lamins are among the most long-lived proteins in cells[27]. Besides *RAS*-induced senescence, we found that Atg7 inhibition also attenuates lamin B1 loss triggered by oxidative stress and DNA damage-induced senescence (Extended Data Fig. 6c–e). Further, mCherry–GFP–lamin B1 expressed in Atg7 knockdown HRasV12 cells displayed normal induction of nuclear membrane blebs but deficient cytoplasmic mCherry signals (Extended Data Fig. 6f, g). These data suggest that inhibition of autophagy leads to a profound defect in the nucleus-to-cytoplasm transport of lamin B1.

Lamin B1 plays an important role in cell proliferation and senescence[17]. Forced knockdown of lamin B1 causes premature senescence[16,17], whereas overexpression of lamin B1 delays senescence[17]. Restoration of lamin B1 in already-established senescent cells is not sufficient to revert senescence *in vitro* (Extended Data Fig. 6h, i). Consistent with the compromised lamin B1 degradation, we found that Atg7 knockdown cells showed delayed HRasV12-induced senescence, as judged by reduced levels of p16 (Fig. 4a and Extended Data Fig. 6j) and delayed induction of senescence-associated β-galactosidase (β-gal) (Extended Data Fig. 6k).

We mapped the LC3–lamin B1 interaction and discovered that LC3 R10 and R11 are essential for lamin B1 binding, from *in vitro* pull-down, *in vivo* co-IP, BiFC, and ChIP experiments (Fig. 4b and Extended Data Fig. 7a–f). Moreover, while LC3-wild type (WT) showed co-localization with CCF, the LC3 mutant failed to do so (Extended Data Fig. 7g). On the lamin B1 end, the region between Coil 2 and the immunoglobulin (Ig)-fold of lamin B1 is necessary for LC3 binding (Fig. 4c and Extended Data Fig. 8a–c). Notably, this region

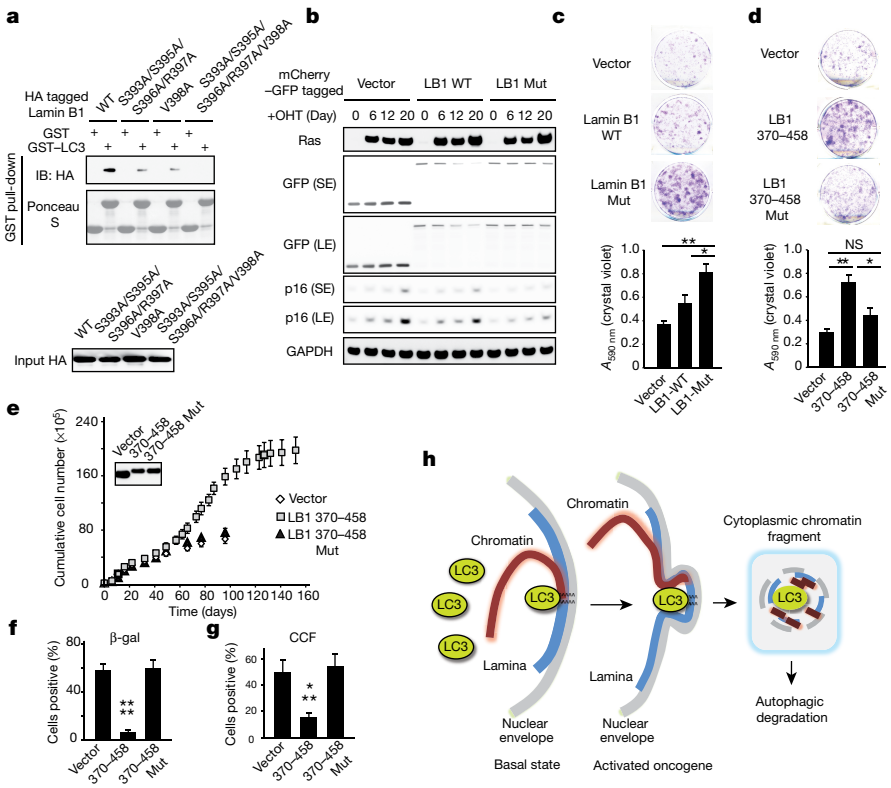


Figure 5 | LC3–lamin B1 interaction is required for lamin B1 degradation and cellular senescence. a, *In vitro* translated proteins were subjected to GST–LC3B pull-down. **b**, BJ ER:HRasV12 cells were analysed by immunoblotting. Uncropped blots are in Supplementary Figure 1. **c**, **d**, Colony formation analysis of BJ ER:HRasV12 cells. $A_{590\,nm}$, absorbance at 590 nm. **e**, Mid-life BJ fibroblasts stably expressing mCherry–GFP-tagged constructs were recorded for growth. Uncropped blots are in Supplementary Figure 1. **f**, Day 60, quantified for β-gal positivity. **g**, Day 101, quantified for cytoplasmic DAPI. Bars, mean ± s.e.m. (**c**, **d**), s.d. (**f**, **g**); $n = 3$ (**c**, **d**), $n = 4$ (**f**, **g**); $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$, $^{****}P < 0.0001$; NS, non-significant; one-way ANOVA coupled with Tukey's *post hoc* test. **h**, Schematic illustration of autophagy degradation of nuclear lamina.

(390–438) is the most evolutionarily conserved domain among all vertebrate lamin B1 (Extended Data Fig. 8d, e). The region, along with 20 amino-acid flanking sequence at the amino and carboxy (N and C) termini (resulting in the fragment 370–458), is sufficient to bind LC3 (Fig. 4c, d and Extended Data Fig. 8f), while the homologous regions on other lamins fail to bind LC3 (Fig. 4d). Examination of the amino-acid sequences revealed that lamins A/C harbour several distinct residues compared with lamin B1, and that lamin B2 has two insertions in the region (Extended Data Fig. 8d), which possibly alters the proper peptide folding for LC3 interaction.

The 370–458 region of lamin B1 contains its nuclear localization signal (NLS) (Fig. 4c), hence the fragment localizes to the nucleus (Extended Data Fig. 8g) and is able to interact with endogenous LC3 (Fig. 4e). Overexpression of this fragment decreases endogenous LC3–lamin B1 interaction, but does not affect LC3 lipidation, LC3 binding to p62 (Fig. 4e), or p62 degradation upon starvation (Extended Data Fig. 8h). When expressed in HRasV12 cells, the fragment impairs lamin B1 downregulation, accompanied by an attenuated senescence (Fig. 4f and Extended Data Fig. 8i–k).

We further identified the essential residues within lamin B1 for binding to LC3, and found that simultaneously substituting the residues S393, S395, S396, R397, and V398 to alanine abrogates the interaction with LC3 (Fig. 5a and Extended Data Fig. 9a–g). In control cells, this lamin B1 substitution mutant shows a normal nuclear peripheral pattern (Extended Data Fig. 9h) and is able to interact with endogenous lamin A and lamin B1 (Extended Data Fig. 9j). However, in HRasV12 cells, the mutant showed attenuated protein downregulation compared with WT lamin B1 (Fig. 5b and Extended Data Fig. 9k), and dramatically reduced cytoplasmic lamin B1 signals (Extended Data Fig. 9h, i), indicating that the mutant has a profound deficiency in nucleus-to-cytoplasm transport. Consequently, the lamin B1 mutant-expressing cells delayed HRasV12-induced senescence with a higher efficiency than WT lamin B1 (Fig. 5b and Extended Data Fig. 9l), and significantly promoted the growth of colonies in colony-formation analysis (Fig. 5c). Furthermore, we used our lamin B1 370–458 peptide that blocks the LC3–lamin B1 interaction and inhibits senescence (Fig. 4e, f). Introducing point mutations as mapped above (Fig. 5a) abrogates the peptide association with LC3 (Extended Data Fig. 10a). While the 370–458 peptide delayed cellular senescence induced by HRasV12, the 370–458 mutant failed to do so (Fig. 5d and Extended Data Fig. 10b). Besides oncogene-induced senescence, the peptide also significantly delayed replicative senescence and the appearance of CCF (Fig. 5e–g and Extended Data Fig. 10c–e). Taken together, these data indicate that the LC3–lamin B1 interaction plays an essential role in reinforcing cellular senescence, which both suppresses oncogene activity and limits cellular lifespan.

In this study, we discovered lamin B1 as a selective mammalian autophagy substrate upon oncogenic and genotoxic insults (illustrated in Fig. 5h). Recently, starvation-induced nuclear autophagy was discovered in yeast[28], which is devoid of nuclear lamina and malignancies. In contrast, we show that mammalian lamin B1 degradation does not occur during starvation. Recent studies reveal that downregulation of lamin B1 impairs cell proliferation and DNA repair[16,17,29,30], and leads to large-scale alterations in chromatin[16]. These dramatic changes are unlikely to happen during starvation, but are probably beneficial in restraining oncogenic and tumorigenic insults. Our study suggests that LC3–lamin B1 interaction occurs in the basal cellular state, and, upon aberrant cellular activities, initiates lamin B1 degradation (Fig. 5h) thus driving senescence to restrain cell proliferation. Hence, selective nuclear lamina degradation by autophagy may play a role in restricting tumorigenesis and maintaining cell and tissue integrity.

Although our current work focuses on lamin B1, we anticipate that other nuclear substrates of autophagy have roles in tumour suppression

and other physiological/pathological scenarios. This study establishes a new perspective in understanding mammalian autophagy—from the nucleus.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 16 July; accepted 4 September 2015.
Published online 28 October 2015.

1. Levine, B. & Kroemer, G. Autophagy in the pathogenesis of disease. *Cell* **132,** 27–42 (2008).
2. Mizushima, N., Levine, B., Cuervo, A. M. & Klionsky, D. J. Autophagy fights disease through cellular self-digestion. *Nature* **451,** 1069–1075 (2008).
3. Choi, A. M., Ryter, S. W. & Levine, B. Autophagy in human health and disease. *N. Engl. J. Med.* **368,** 651–662 (2013).
4. Kabeya, Y. *et al.* LC3, a mammalian homologue of yeast Apg8p, is localized in autophagosome membranes after processing. *EMBO J.* **19,** 5720–5728 (2000).
5. Mizushima, N., Yoshimori, T. & Ohsumi, Y. The role of Atg proteins in autophagosome formation. *Annu. Rev. Cell Dev. Biol.* **27,** 107–132 (2011).
6. Rogov, V., Dötsch, V., Johansen, T. & Kirkin, V. Interactions between autophagy receptors and ubiquitin-like proteins form the molecular basis for selective autophagy. *Mol. Cell* **53,** 167–178 (2014).
7. Drake, K. R., Kang, M. & Kenworthy, A. K. Nucleocytoplasmic distribution and dynamics of the autophagosome marker EGFP-LC3. *PLoS One* **5,** e9806 (2010).
8. Huang, R. *et al.* Deacetylation of nuclear LC3 drives autophagy initiation under starvation. *Mol. Cell* **57,** 456–466 (2015).
9. Simon, H. U., Yousefi, S., Schmid, I. & Friis, R. ATG5 can regulate p53 expression and activation. *Cell Death Dis.* **5,** e1339 (2014).
10. Lee, I. H. *et al.* Atg7 modulates p53 activity to regulate cell cycle and survival during metabolic stress. *Science* **336,** 225–228 (2012).
11. Shimi, T. *et al.* The A- and B-type nuclear lamin networks: microdomains involved in chromatin organization and transcription. *Genes Dev.* **22,** 3409–3421 (2008).
12. Kerppola, T. K. Bimolecular fluorescence complementation(BiFC) analysis as a probe of protein interactions in living cells. *Annu. Rev. Biophys.* **37,** 465–487 (2008).
13. Guelen, L. *et al.* Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature* **453,** 948–951 (2008).
14. Lund, E., Oldenburg, A. R. & Collas, P. Enriched domain detector: a program for detection of wide genomic enrichment domains robust against local variations. *Nucleic Acids Res.* **42,** e92 (2014).
15. Sadaie, M. *et al.* Redistribution of the Lamin B1 genomic binding profile affects rearrangement of heterochromatic domains and SAHF formation during senescence. *Genes Dev.* **27,** 1800–1808 (2013).
16. Shah, P. P. *et al.* Lamin B1 depletion in senescent cells triggers large-scale changes in gene expression and the chromatin landscape. *Genes Dev.* **27,** 1787–1799 (2013).
17. Shimi, T. *et al.* The role of nuclear lamin B1 in cell proliferation and senescence. *Genes Dev.* **25,** 2579–2593 (2011).
18. Freund, A., Laberge, R. M., Demaria, M. & Campisi, J. Lamin B1 loss is a senescence-associated biomarker. *Mol. Biol. Cell* **23,** 2066–2075 (2012).
19. Ivanov, A. *et al.* Lysosome-mediated processing of chromatin in senescence. *J. Cell Biol.* **202,** 129–143 (2013).
20. Serrano, M., Lin, A. W., McCurrach, M. E., Beach, D. & Lowe, S. W. Oncogenic ras provokes premature cell senescence associated with accumulation of p53 and p16INK4a. *Cell* **88,** 593–602 (1997).
21. Collado, M., Blasco, M. A. & Serrano, M. Cellular senescence in cancer and aging. *Cell* **130,** 223–233 (2007).
22. Young, A. R. *et al.* Autophagy mediates the mitotic senescence transition. *Genes Dev.* **23,** 798–803 (2009).
23. Liu, H. *et al.* Down-regulation of autophagy-related protein 5(ATG5) contributes to the pathogenesis of early-stage cutaneous melanoma. *Sci. Transl. Med.* **5,** 202ra123 (2013).
24. Roberts, P. *et al.* Piecemeal microautophagy of nucleus in *Saccharomyces cerevisiae*. *Mol. Biol. Cell* **14,** 129–141 (2003).
25. Mizushima, N., Yoshimori, T. & Levine, B. Methods in mammalian autophagy research. *Cell* **140,** 313–326 (2010).
26. Pankiv, S. *et al.* p62/SQSTM1 binds directly to Atg8/LC3 to facilitate degradation of ubiquitinated protein aggregates by autophagy. *J. Biol. Chem.* **282,** 24131–24145 (2007).
27. Toyama, B. H. *et al.* Identification of long-lived proteins reveals exceptional stability of essential cellular structures. *Cell* **154,** 971–982 (2013).
28. Mochida, K. *et al.* Receptor-mediated selective autophagy degrades the endoplasmic reticulum and the nucleus. *Nature* **522,** 359–362 (2015).
29. Butin-Israeli, V. *et al.* Role of lamin B1 in chromatin instability. *Mol. Cell. Biol.* **35,** 884–898 (2015).
30. Dreesen, O. *et al.* Lamin B1 fluctuations have differential effects on cellular proliferation and senescence. *J. Cell Biol.* **200,** 605–617 (2013).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank members of the Berger, Adams, and Goldman laboratories for technical assistance and discussions. We acknowledge A. L. Stout for help with confocal microscopy, and the electron microscopy resource laboratory for assistance on TEM. We thank Z. Yue for sharing the GFP antibody and reading the manuscript, and M. Narita and R. Salama for help with LADs definition. Z.D. is supported by a fellow award from the Leukemia & Lymphoma Society. B.C.C. is supported by career development awards from the Dermatology Foundation, Melanoma Research Foundation, and American Skin Association. S.L.B., P.D.A. and R.M. are supported by NIA P01 grant (P01AG031862). S.L.B. is also supported by NIH R01 CA078831. R.D.G. is supported by R01 GM106023 and the Progeria Research Foundation.

Author Contributions Z.D., A.I., P.D.A., and S.L.B. conceived the project. Z.D. performed most of the experiments. C.X., G.D., B.C.C., A.M.D., and P.P.S. performed and analysed ChIP-seq. T.S. performed three-dimensional structural illumination microscopy imaging. T.S., S.A.A., and R.D.G. contributed novel lamin reagents and experimental design. J.-A.P., J.M.C., and W.-X.Z. contributed novel autophagy and senescence reagents. J.Z. performed Atg7 knockdown. M.D.R. and R.M. contributed to the biochemistry characterization of LC3–lamin B1 interaction. T.L. and T.J. contributed novel autophagy constructs and experimental design. Z.D., P.D.A., and S.L.B. composed the manuscript. All authors reviewed the manuscript and discussed the work.

Author Information LC3 and lamin B1 ChIP-seq data have been deposited in the NCBI Gene Expression Omnibus (GEO) database under accession number GSE63440. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to S.L.B. (bergers@upenn.edu).

METHODS

Data reporting. No statistical methods were used to predetermine sample size. The experiments were not randomized. The investigators were not blinded to allocation during experiments and outcome assessment.

Cell lines and culture. IMR90, mouse embryonic fibroblasts, and HEK293T were described previously[16,31]. Primary BJ fibroblasts were purchased from ATCC. Cell line identities were not further authenticated. The cells were cultured in DMEM supplemented with 10% fetal bovine serum (FBS), 100 U ml^{-1} penicillin, and 100 μg ml^{-1} streptomycin (Invitrogen), and were intermittently tested for mycoplasma. IMR90 and BJ were cultured under physiological oxygen (3%), except for the H_2O_2 treatment, in which cells were cultured in an incubator with 20% oxygen, and the experiments involved in live-cell imaging. For primary cell cultures, cells were briefly washed with PBS, trypsinized at 37 °C for 2–4 min, and passaged at no more than 1:4 dilutions. Cells were counted with a Countess automated cell counter (Life Technologies), and the numbers were recorded where growth curves were generated. HEK293T cells were transfected using Lipofectamine 2000 (Invitrogen). For amino-acid starvation, cells were incubated in Hank's buffer (with calcium and glucose) supplemented with 10% dialysed FBS and 1% HEPES (Invitrogen). For amino-acid and serum deprivation, cells were cultured in Hank's buffer plus 1% HEPES.

Retrovirus and lentivirus infection. Stable cell lines were made by retrovirus or lentivirus infection, as previously described[31], with slight modifications. Retroviral constructs were transfected to Phoenix packaging cell line. Lentiviral pLKO constructs were transfected with packaging plasmids to HEK293T cells. Viral supernatant was filtered through a 0.45-μm filter, supplemented with 8 μg ml^{-1} polybrene, and mixed with trypsinized recipient cells. pLNCX-ER:HRasV12, WZL-hygro, and WZL-HRasV12-hygro viral constructs were described elsewhere[20,22]. sh-Atg7 hairpin sequence GGAGTCACAGCTCTTCCTTAC was from ref. 22, and cloned into Tet-pLKO-puro 'all-in-one' tetracycline-inducible vector[32]. Doxycyclin 100 ng ml^{-1} was added to IMR90 to induce knockdown of Atg7. Another pLKO-shAtg7 construct (TRCN0000007587) was purchased from Sigma-Aldrich and used in BJ fibroblasts. The infected cells were selected with puromycin, neomycin, or hygromycin for about 1 week.

Reagents and antibodies. Rapamycin was purchased from Millipore. H_2O_2 was from Fisher Scientific. 4-Hydroxytamoxifen and etoposide was from Sigma-Aldrich. The following antibodies were used: LC3 (MBL PM036 for WB of mouse embryonic fibroblasts; Cell Signaling Technology 3868 for immunoprecipitation, ChIP, IF, WB; Cell Signaling Technology 2775 for WB), β-tubulin (Sigma-Aldrich T4026), calreticulin (Cell Signaling Technology 12238), COX IV (Cell Signaling Technology 4850), Atg5 (Cell Signaling Technology 8540), Atg7 (Cell Signaling Technology 8558), lamin B1 (Abcam ab16048), lamin B2 (Abcam ab8983), lamins A/C (Millipore MAB3211), GFP (Roche 11 814 460 001 and Abcam ab290), p62 (Abnova H00008878-M01), GAPDH (Fitzgerald Industries 10R-G109A), p16 (Abcam ab16123), Ras (Millipore 05-516), HA (Sigma-Aldrich H3663), H3K27me3 (Active Motif 39538), H3K9me3 (Abcam ab8898), LAMP1 (Iowa Hybridoma Bank H4a3-s), and Flag (Sigma-Aldrich F1804).

Plasmids. GST, GST–LC3A, B, C, and GST–LC3B mutants/truncations were described elsewhere[33]. GFP, HA/Flag/GFP–LC3 WT and mutants, GFP–Beclin 1, GFP–ULK1, GFP–lamin B1, and split Venus constructs were described previously[17,31,34,35]. pBabe–mCherry–GFP–LC3 (ref. 36) was purchased from Addgene, and LC3 was truncated to make pBabe–mCherry–GFP, and then lamin B1 sequences were cloned. Lamin B1 truncations/mutations were made from pEGFP–lamin B1 for direct transfection, pBabe–mCherry–GFP–lamin B1 for retrovirus, or pT7-NHA-lamin B1 for *in vitro* translation. Tet-inducible lentiviral GFP–lamin B1 was made by cloning the GFP–lamin B1 fragment into pTRIPZ. All new constructs in this study were verified by DNA sequencing.

Western blotting. Cells were lysed in buffer containing 50 mM Tris pH 7.5, 0.5 mM EDTA, 150 mM NaCl, 1% NP40, 1% SDS, supplemented with 1:100 Halt Protease inhibitor cocktail (Thermo Scientific). The lysates were briefly sonicated, and supernatants were subjected to electrophoresis using NuPAGE Bis-Tris precast gels (Life Technologies). After transferring to nitrocellulose membrane, 5% milk in TBS supplemented with 0.1% Tween 20 (TBST) was used to block the membrane at room temperature (~25 °C) for 1 h. Primary antibodies were diluted in 5% BSA in TBST, and incubated at 4 °C overnight. The membrane was washed three times with TBST, each for 10 min, followed by incubation of HRP-conjugated secondary antibodies at room temperature for 1 h, in 5% milk/TBST. The membrane was washed again three times, and imaged by a Fujifilm LAS-4000 imager.

Immunoprecipitation. Cells were lysed in immunoprecipitation buffer containing 20 mM Tris, pH 7.5, 137 mM NaCl, 1 mM $MgCl_2$, 1 mM $CaCl_2$, 1% NP-40, 10% glycerol, supplemented with 1:100 Halt protease and phosphatase inhibitor cocktail (Thermo Scientific) and benzonase (Novagen) at 12.5 U ml^{-1}. Benzonase is essential to release chromatin-bound proteins to supernatant, and $MgCl_2$ is critical for its activity. The lysates were rotated at 4 °C for 30–60 min. The supernatant was incubated with antibody-conjugated Dynabeads (Life Technologies), and rotated at 4 °C overnight. The immunoprecipitation was washed and collected by magnet, for five times with immunoprecipitation buffer, and boiled with NuPAGE loading dye. Samples were analysed by western blotting.

***In vitro* translation.** Cell-free *in vitro* translation was performed using the 1-Step *In Vitro* Translation Kit (Thermo Scientific), following the manufacturer's guidance. Target proteins were cloned into pT7CFE1-NHA vector (with N-terminal HA tag) and translated *in vitro* at 30 °C.

Bacteria expression and GST pull-down. GST-tagged constructs were transformed into BL21-CodonPlus *Escherichia coli* and purified with glutathione beads (Life Technologies). Lamin B1 370–458 and 390–438 fragments were cloned into GST construct with a TEV protease recognition site between GST and the cloned sequences. The expressed proteins were loaded and purified with glutathione agarose beads, and digested with His-tagged TEV protease. The resulting supernatant was further purified with Ni-NTA beads (Qiagen) to remove His-tagged TEV protease.

For GST pull-down, bacterial lysates were incubated with glutathione beads at 4 °C for 2 h and washed four times with buffer containing 50 mM Tris, pH 7.5, 150 mM NaCl, 1% Triton X-100, 1 mM DTT, supplemented with 100 μM PMSF. The purified proteins or *in vitro* translated proteins were diluted in binding buffer (20 mM Tris, pH 7.5, 137 mM NaCl, 1 mM $MgCl_2$, 1 mM $CaCl_2$, 1% NP-40, supplemented with 1:1,000 Halt Protease inhibitor cocktail) and then pre-cleared with GST at 4 °C for 1 h. The resulting supernatant was then subjected to GST pull-down with GST or GST fusion proteins. The product was washed four times with binding buffer and boiled with NuPAGE loading dye for immunoblotting analysis. Purified lamin B1 protein was purchased from Origene.

Immunofluorescence and live-cell imaging. For immunofluorescence, cells were fixed in 4% paraformaldehyde in PBS for 30 min at room temperature. Cells were washed twice with PBS, and permeabilized with 0.5% Triton X-100 in PBS for 10 min. After washing two times, cells were blocked in 10% BSA in PBS for 1 h at room temperature. Cells were incubated with primary antibodies in 5% BSA in PBS supplemented with 0.1% Tween 20 (PBST) overnight at 4 °C. The next day, cells were washed four times with PBST, each for 10 min, followed by incubation with Alexa Fluor-conjugated secondary antibody (Life Technologies) in 5% BSA/PBST for 1 h at room temperature. Cells were then washed four times in PBST, incubated with 1 μg ml^{-1} DAPI in PBS for 5 min, and washed twice with PBS. The slides were then mounted with ProLong Gold (Life Technologies) and imaged with a Leica TCS SP8 fluorescent confocal microscope. The slides were mounted with ProLong Diamond (Life Technologies) for 5 days at room temperature for super-resolution microscopy.

Three-dimensional structural illumination microscopy was performed using N-SIM Super-resolution Microscope System (Nikon) with an oil immersion objective lens CFI SR (Apochromat TIRF ×100, 1.49 numerical aperture; Nikon). Twenty to forty-one optical sections were collected with a 200 nm interval between neighbouring sections.

For live-cell imaging, mCherry–GFP–lamin B1 HRasV12 cells were plated onto a 35 mM glass bottom dish (MatTek P35G-0-14-C) pre-coated with poly-L-lysine (Sigma-Aldrich). The dish was imaged with a spinning disk fluorescent confocal microscope (Olympus IX71 and IX81 Inverted System, coupled with an Andor iXon3 EMCCD camera, with motorized *x*–*y* stage, Okolab stagetop incubation chamber, and MetaMorph acquisition software). Cells were imaged overnight every 15 min. Twelve *z*-sections were acquired covering the entire individual cell. Images were viewed and presented as the maximum projection from all *z*-sections.

TEM. For immuno-gold TEM, GFP–lamin B1 expressing IMR90 cells were subjected to high-pressure freezing. The samples were then dehydrated by freeze substitution methods for 72 h at −90 °C in 0.1% uranyl acetate/acetone followed by embedding in Lowicryl HM20 at −50 °C with 360 nm light polymerization of the resin for 48 h. Resin-embedded cells were sectioned at 70 nm thickness. GFP–lamin B1 was detected with a GFP antibody[35] diluted 1:50 in 5% BSA, 0.1% fish gelatin, in PBS. Gold colloids (10 nm) conjugated to goat anti-rabbit (Electron Microscopy Sciences) at 1:200 was used for secondary detection of GFP–antigen conjugates followed by a 0.2% gluteraldehyde post-fix to stabilize the immuno-protein complexes. Imaging was performed at 80 keV on a JEOL 1010 at indicated magnifications and collected digitally on an AMT side-entry CCD (charge-coupled device) without post-labelling heavy-metal staining. For TEM analysis of ultrastructures of control and HRasV12 IMR90, cells were subjected to high-pressure freezing, followed by standard TEM procedures.

ChIP, RT-qPCR, and ChIP sequencing. These assays were performed as described previously[16] with slight modification. In brief, cells were crosslinked with 1% formaldehyde diluted in PBS, without the addition of other co-crosslinkers, for 5 min at room temperature. After glycin quenching, the cell pellets were lysed in buffer containing 50 mM Tris, pH 7.5, 150 mM NaCl, 1% Triton X-100, 0.1% Na-deoxycholate, 0.1% SDS, supplemented with complete protease inhibitor

cocktail (Thermo Scientific), and sonicated with a Covaris sonicator, resulting in chromatin fragments with an average size of 250 base pairs. The supernatant was diluted ten times with the above buffer without SDS, and subjected to immunoprecipitations with 2 µg of antibody or control IgG conjugated with Dynabeads Protein A or G (Invitrogen) at 4 °C overnight. The beads were then washed five times with buffer containing 50 mM Tris, pH 7.5, 150 mM NaCl, 1% Triton X-100, and once with final wash buffer (50 mM Tris, pH 8.0, 10 mM EDTA, 50 mM NaCl), followed by elution with incubation of elution buffer (final wash buffer plus 1% SDS) at 65 °C for 30 min with agitation in a thermomixer. The ChIP and input were then purified and used for qPCR analysis or for constructing sequencing libraries with a NEBNext Ultra kit (New England Biolabs). For ChIP-sequencing, the libraries were quantified (Kapa Biosystems) and were single-end sequenced on an Illumina NextSeq 2000.

The following primers were used for qPCR analyses of LADs. LAD1: forward, AGAGACGTGGCGTGTGTCC; reverse, GGCACTGAAGCCACCTCTGT (chromosome 4: 190524973–190525023). LAD2: forward, ATTTGC ACAATCTGAGGGCG; reverse, CTGGGCAATTCCCTTGGTAGT (chromosome 7: 35434121–35434171). LAD3: forward, GCATCCATTTCACATCCTTGG; reverse, CCCATTGCCTCTGAAGTTTTGT (chromosome 8: 130184820–130184870).

Subcellular fractionation. This was performed with the subcellular fractionation kit for cultured cell (Thermo Scientific 78840) according to the manufacturer instructions, with slight modification. Benzonase (Novagen) was used to digest chromatin-bound proteins in the nuclear fraction, in the buffer supplemented with 5 mM $MgCl_2$.

Senescence-associated β-gal assay. β-Galactosidase assays were performed using a cellular senescence assay kit (Chemicon KAA002), according to the manufacturer's protocol. Cells were incubated with β-gal detection solution at 37 °C overnight, and quantified under regular light microscopy. At least 200 cells were scored for β-gal positivity with over four different fields.

Computational methods. Alignment of vertebrate lamin B1 proteins was done using ClustalX 2.1 (ref. 37). Computational analysis of ChIP-seq was performed as previously described and as follows.

Data source: H3 (GEO accession number GSM897555), H3K4me3 (GEO GSM897556). H3K9me3 ChIP-seq data (GEO GSM942075 and GSM942119) were published elsewhere[15,38]. LC3 and lamin B1 ChIP-seq data in this study have been deposited in the GEO (http://www.ncbi.nlm.nih.gov/geo) under accession number GSE63440.

Alignment of lamin B1, LC3, and input: all ChIP-seq data were aligned to the GRCh37 (hg19) assembly of the human genome using bowtie2 with command-line parameters -k1 -N1 —local (allowing and reporting a single alignment per read with one or zero mismatch permitted in the seed region).

Track generation: ChIP-seq visualization tracks were created in the following way. Aligned sequence tags were subjected to BEDTools' genomeCoverageBed tool, making bedGraphs that were multiplied by the RPM coefficient. A similarly normalized input bedGraph was then subtracted from lamins B1 and LC3, and bedGraphs were made into bigWigs using the University of California Santa Cruz Genome Browser's bedGraphToBigWig utility.

Box plot: aligned tag counts were assessed for each LAD for all marks under study, as well as the corresponding input and H3. The distribution of ChIP enrichment (ChIP-background) was computed over all LADs or over an equal number of size-matched background regions, sampled from all genomic positions that did not overlap with LADs. Hypothesis testing was done by Mann–Whitney/Wilcoxon tests.

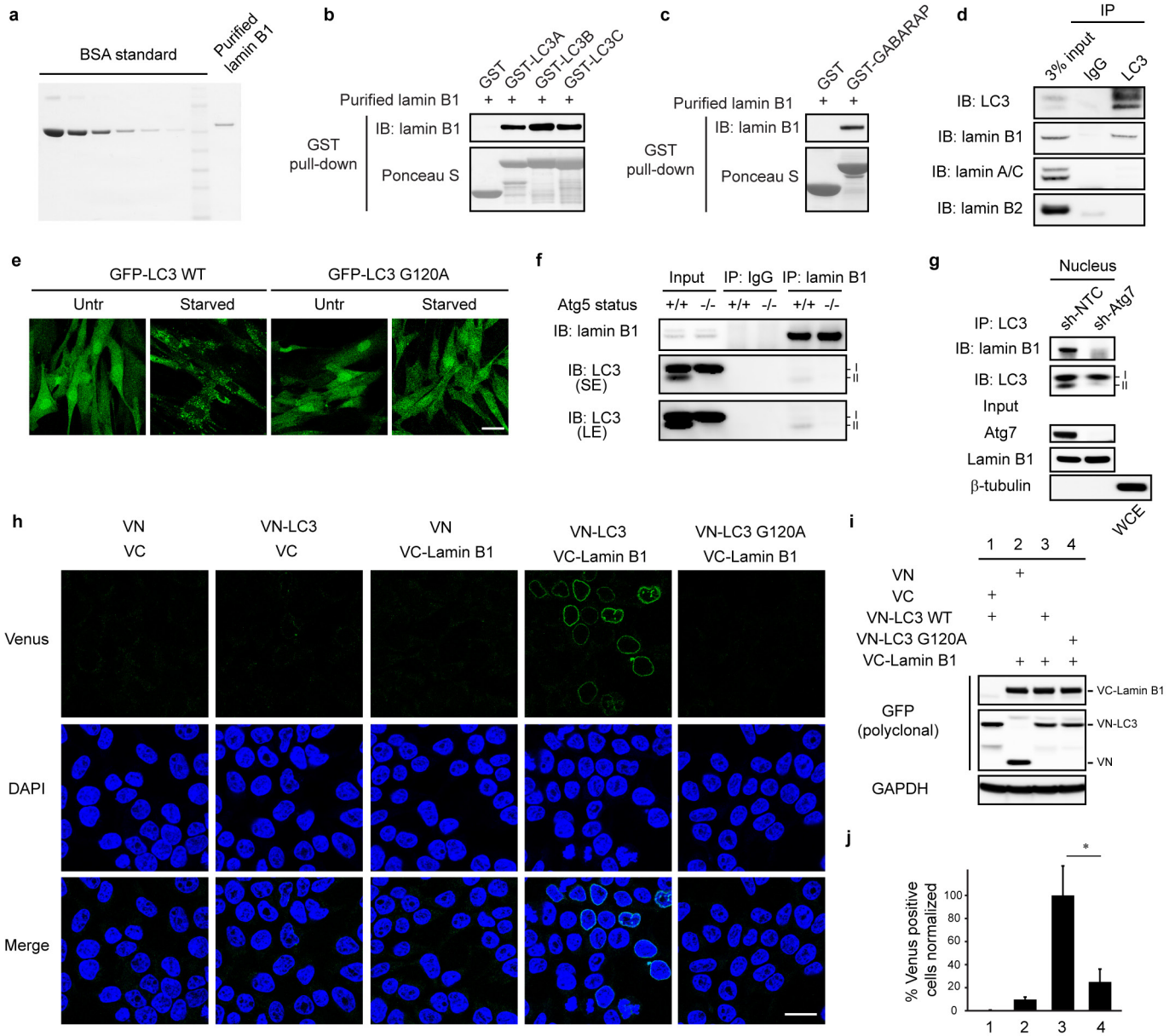
Overlap permutation test: to determine whether LADs were significantly associated with LC3ADs, the number of base pairs in common between LADs and other domains was tabulated using BEDTools intersect (default two-set comparison). In each of 1,000 iterations, LAD coordinates were randomly shuffled using BEDTools, creating 1,000 sets of equal-sized control regions. Each control set was scored for the number of base pairs in common with LC3ADs, and the frequency with which control sets shared more genomic space with other domains than LADs was taken to be an estimate of the probability that a LAD–LC3AD association was not due to chance.

Area under the curve permutation test: for Fig. 2e, a permutation test for LC3, H3K9me3, and H3K4me3 over LADs was performed. In each of 100 iterations, LADs coordinates were randomly shuffled using BEDTools, creating 100 sets of equal-sized non-LADs control regions. LADs as well as each of the 100 non-LAD control sets were scored for LC3, H3K9me3, and H3K4me3 enrichment, and the number of control sets in which the median score was greater than or equal to the median value of the LAD distribution was tabulated. That frequency was taken to be an estimate of the probability that enrichment over LADs was not due to chance; that is, the probability of the null hypothesis that LADs and non-LADs had the same median enrichment. The *P*-value for H3K9me3 was less than 0.01, and the *P*-value for H3K4me3 was 1. This test was repeated using the 75th percentile value as the test statistic and with the 90th percentile value, with the same result in both cases.

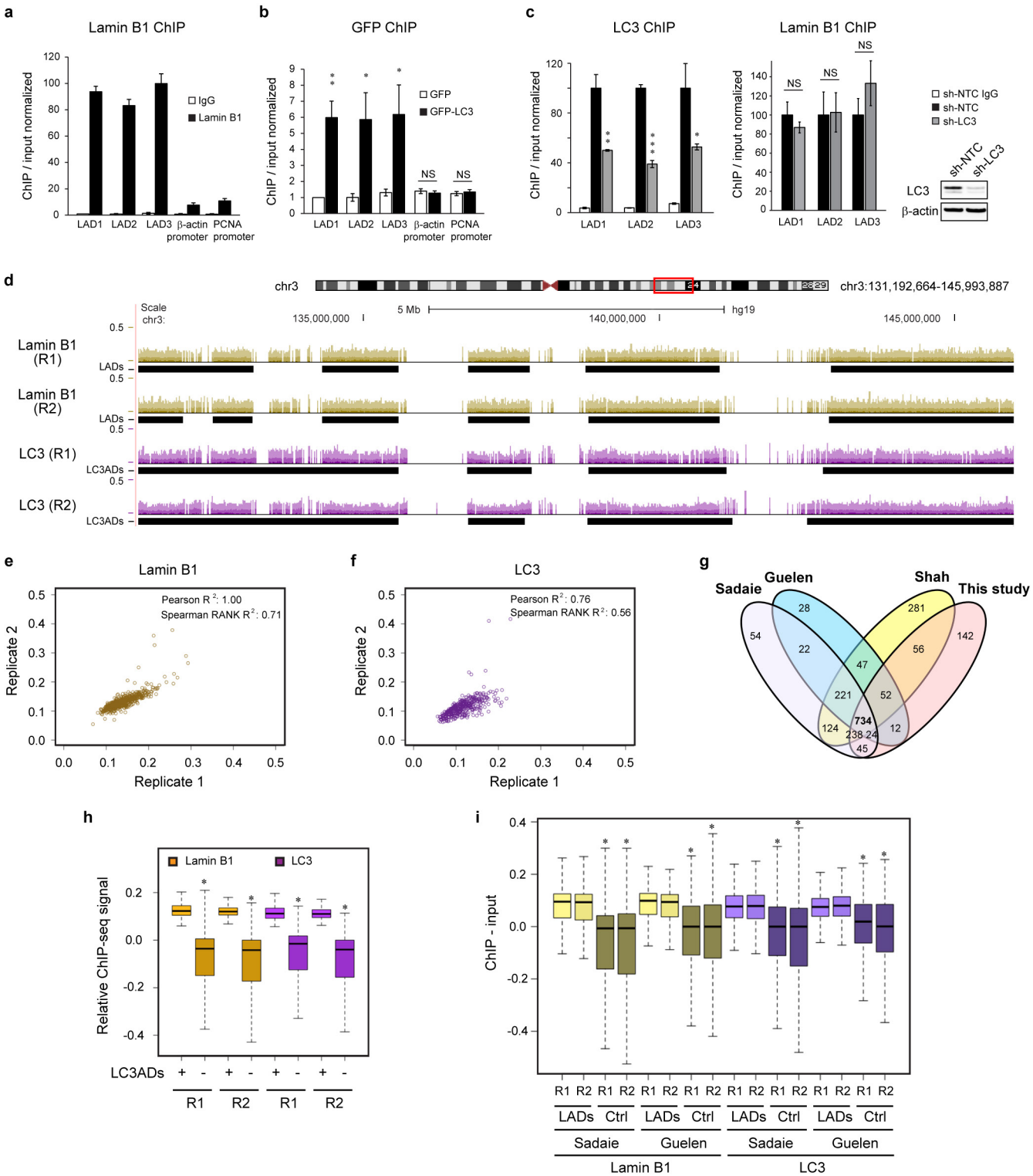
Domain detection: enriched domains for lamin B1 and LC3 were called using EDD[14] with default bin size estimation and gap penalty estimation, and unalignable regions (the hg19 assembly gap track from Genome Reference Consortium) masked. The false discovery rate was controlled at the default value of 5%.

Statistical analysis. Student's *t*-test was used for comparison between two groups. One-way ANOVA coupled with Tukey's *post hoc* test was used for comparisons over two groups. Significance was considered when the P-value was less than 0.05.

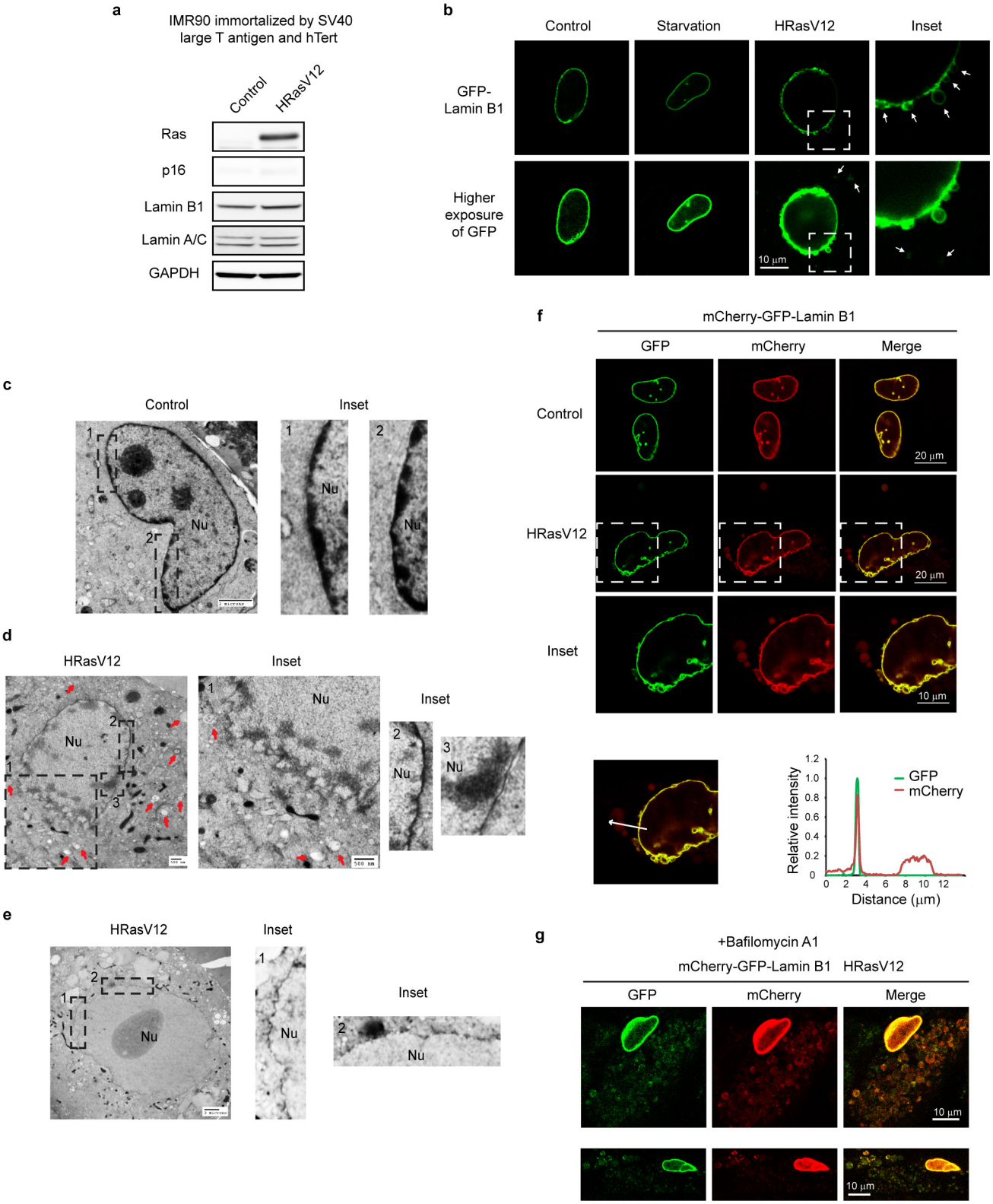
31. Dou, Z. *et al.* Class IA PI3K p110β subunit promotes autophagy through Rab5 small GTPase in response to growth factor limitation. *Mol. Cell* **50,** 29–42 (2013).
32. Wiederschain, D. *et al.* Single-vector inducible lentiviral RNAi system for oncology target validation. *Cell Cycle* **8,** 498–504 (2009).
33. Kirkin, V. *et al.* A role for NBR1 in autophagosomal degradation of ubiquitinated substrates. *Mol. Cell* **33,** 505–516 (2009).
34. Pan, J. A., Ullman, E., Dou, Z. & Zong, W. X. Inhibition of protein degradation induces apoptosis through a microtubule-associated protein 1 light chain 3-mediated activation of caspase-8 at intracellular membranes. *Mol. Cell. Biol.* **31,** 3158–3170 (2011).
35. Zhong, Y. *et al.* Distinct regulation of autophagic activity by Atg14L and Rubicon associated with Beclin 1–phosphatidylinositol-3-kinase complex. *Nature Cell Biol.* **11,** 468–476 (2009).
36. N'Diaye, E. N. *et al.* PLIC proteins or ubiquilins regulate autophagy-dependent cell survival during nutrient starvation. *EMBO Rep.* **10,** 173–179 (2009).
37. Larkin, M. A. *et al.* Clustal W and Clustal X version 2.0. *Bioinformatics* **23,** 2947–2948 (2007).
38. Chandra, T. *et al.* Independence of repressive histone marks and chromatin compaction during senescent heterochromatic layer formation. *Mol. Cell* **47,** 203–214 (2012).



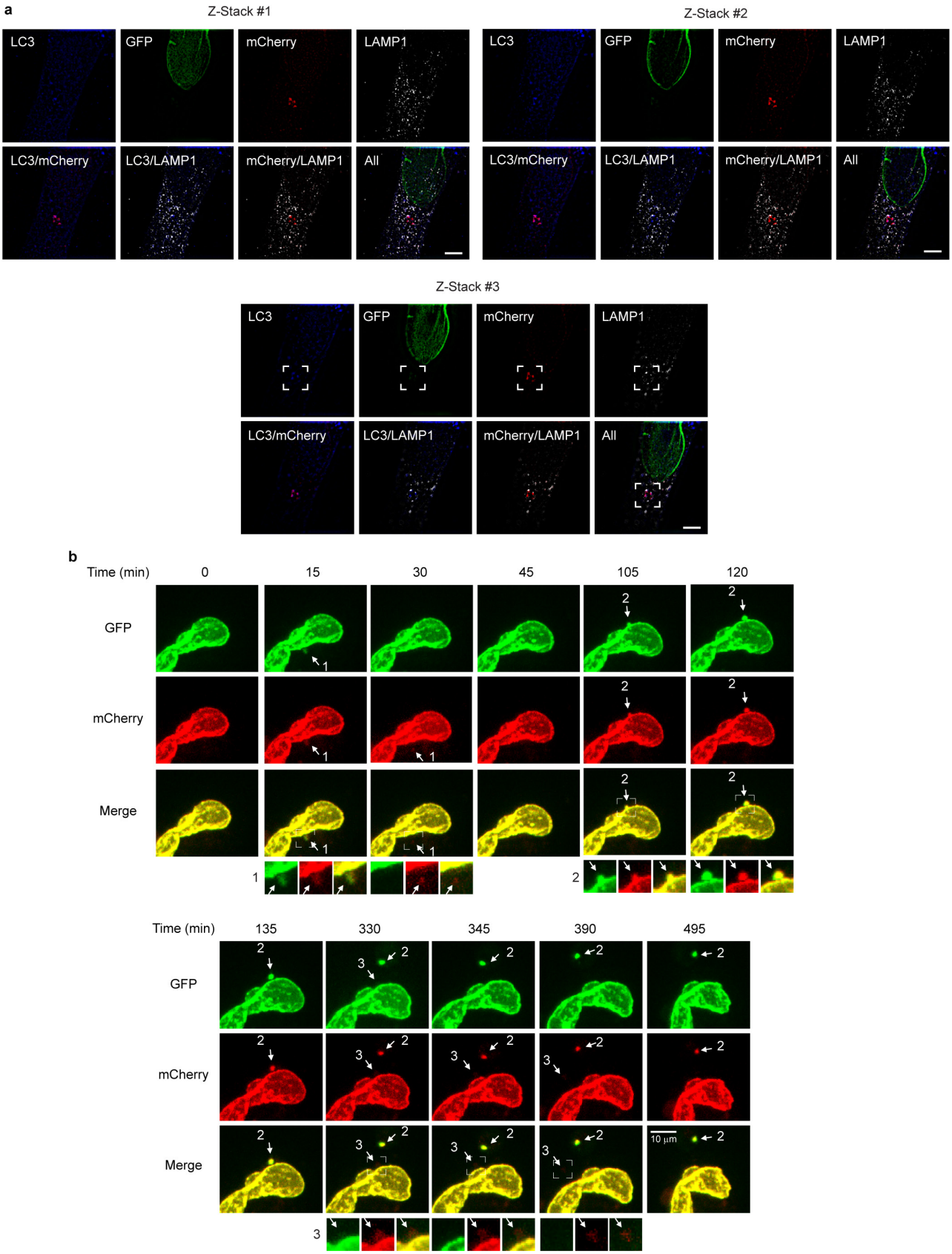
Extended Data Figure 1 | Characterization of LC3 and lamin B1 association. a, Protein gel staining of purified lamin B1 protein. **b**, **c**, Purified lamin B1 protein was subjected to GST pull-down. **d**, Endogenous LC3 immunoprecipitation in HEK293T cells. **e**, IMR90 stably expressing GFP–LC3 constructs were starved and imaged. **f**, Endogenous co-IP in wild-type and Atg5 knockout mouse embryonic fibroblasts. **g**, Nuclear fractions of control and Atg7 knockdown IMR90 cells were analysed by LC3 immunoprecipitation. **h**–**j**, BiFC analysis of LC3–lamin B1 interaction. HeLa cells were transfected with the indicated combination of split Venus constructs and analysed as follows. **h**, Cells were fixed and imaged. **i**, Lysates were analysed by immunoblotting. **j**, Cells were scored for Venus positivity. Bars, mean ± s.d.; $n = 4$, with over 500 cells; $^{*}P < 0.001$; unpaired two-tailed Student's t-test.



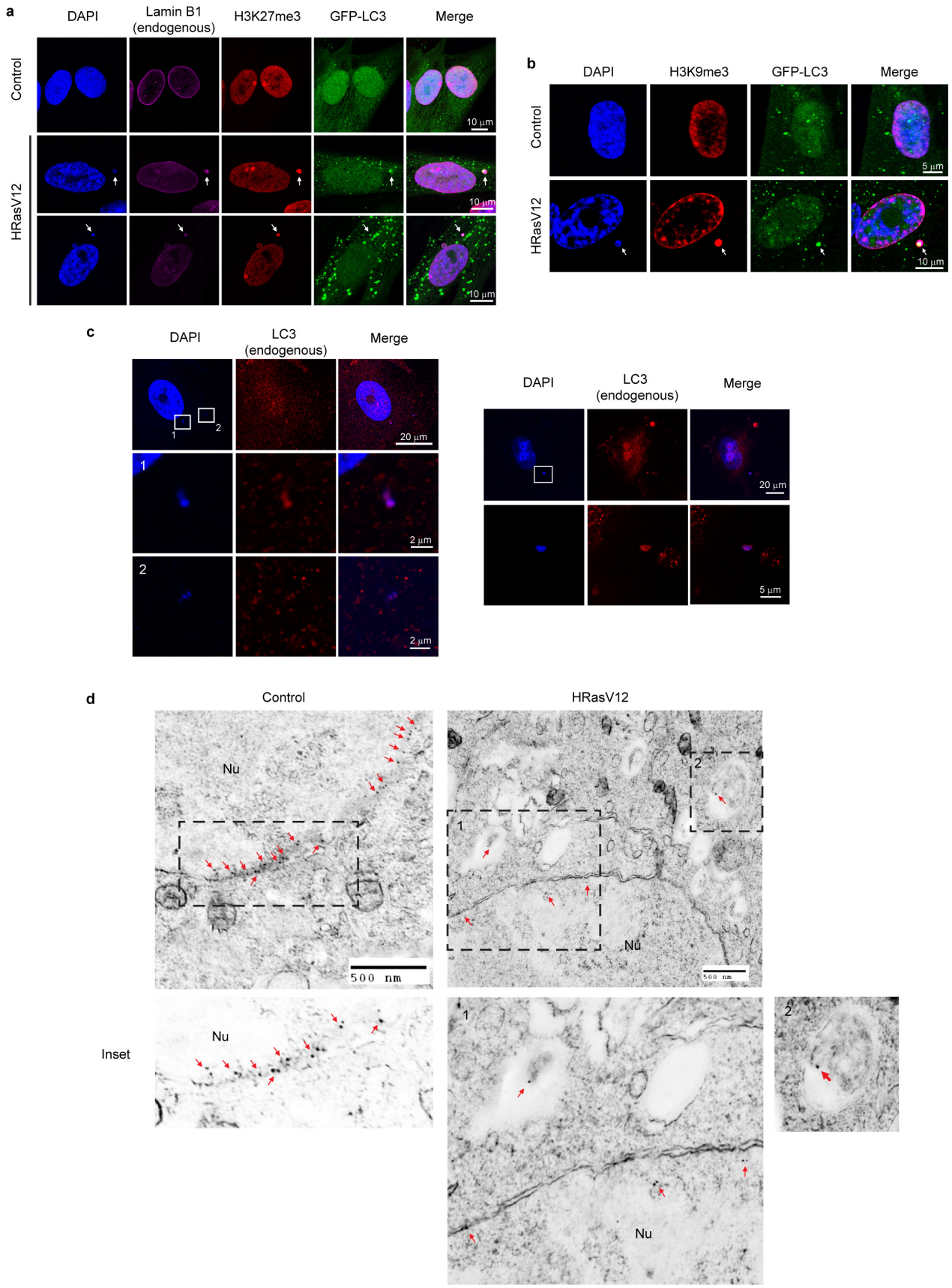
Extended Data Figure 2 | LC3 interacts with LADs on chromatin. **a**, **b**, ChIP–qPCR of proliferating IMR90. **c**, ChIP–qPCR of LC3 knockdown IMR90. Bars, mean ± s.e.m. (**a**, **b**), s.d. (**c**); $n = 3$; $^{*}P < 0.05$, $^{**}P < 0.005$, $^{***}P < 0.0001$; NS, non-significant; unpaired two-tailed Student's *t*-test. **d**–**i**, ChIP-sequencing analyses. **d**, Related to Fig. 2c, a zoom-in window of chromosome 3. **e**, **f**, Analyses of two replicates at LADs and LC3ADs. **g**, Per-nucleotide overlap of published data sets with the LADs called from this study. Number unit: megabases. **h**, Enrichment over LC3ADs. $^{*}P < 2.2 \times 10^{-16}$; one-sided Wilcoxon test. **i**, Analysis of our lamin B1 and LC3 ChIP-seq at LADs defined by other studies, and randomly sampled non-LAD loci (Ctrl). $^{*}P < 2.2 \times 10^{-16}$; one-sided Wilcoxon test.



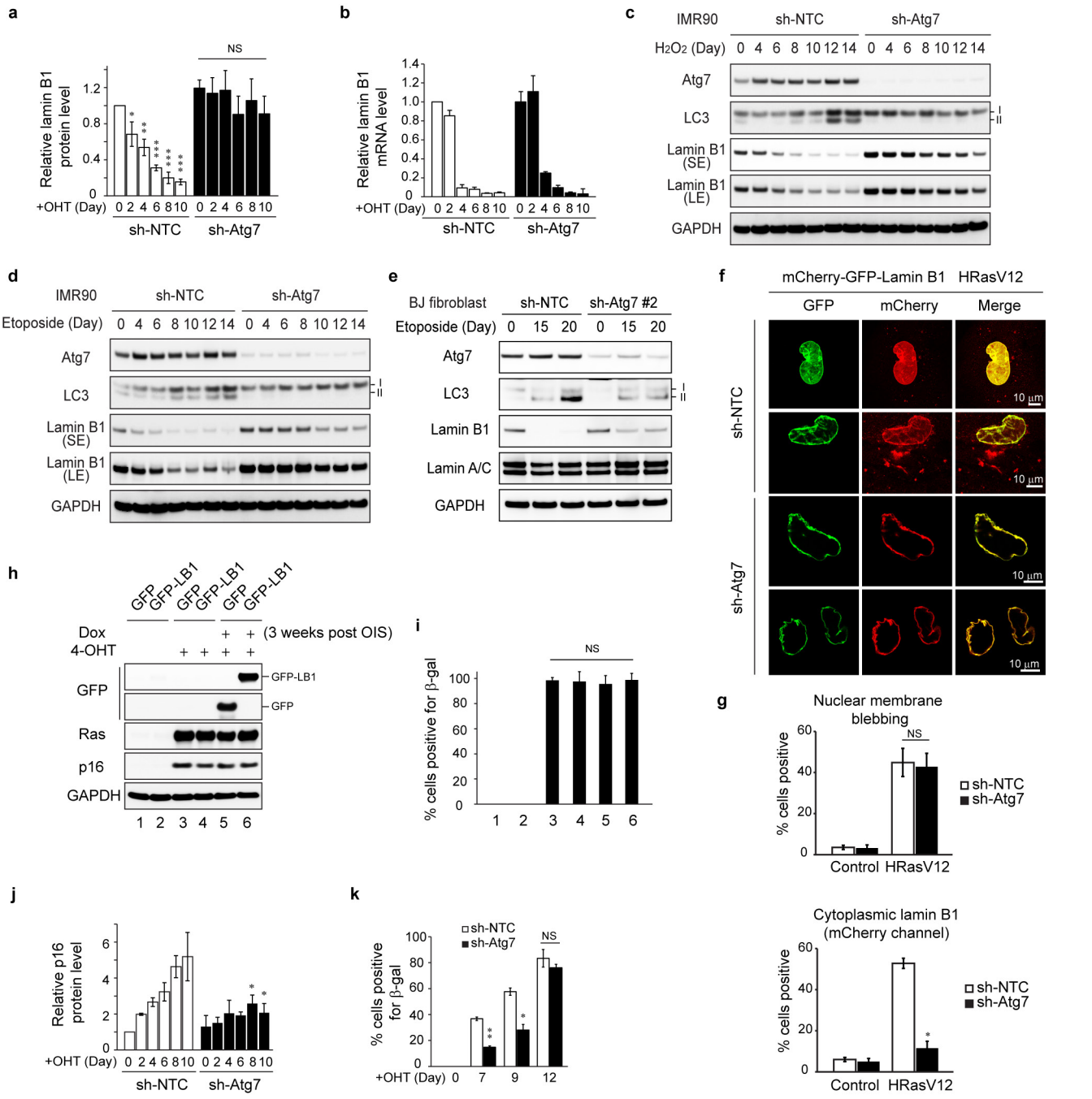
Extended Data Figure 3 | Lamin B1 degradation upon HRasV12-induced senescence. a, Related to Fig. 3b. Immunoblotting of immortalized IMR90. **b**, GFP–lamin B1 stably expressing IMR90 cells were treated as indicated and imaged. Cytoplasmic signals are indicated by arrows. **c**–**e**, TEM analyses of IMR90. Nu, nucleus. **f**, IMR90 cells stably expressing mCherry–GFP–lamin B1 were imaged and quantified. **g**, Cells as in **f** were treated with bafilomycin A1 and imaged under confocal microscopy.



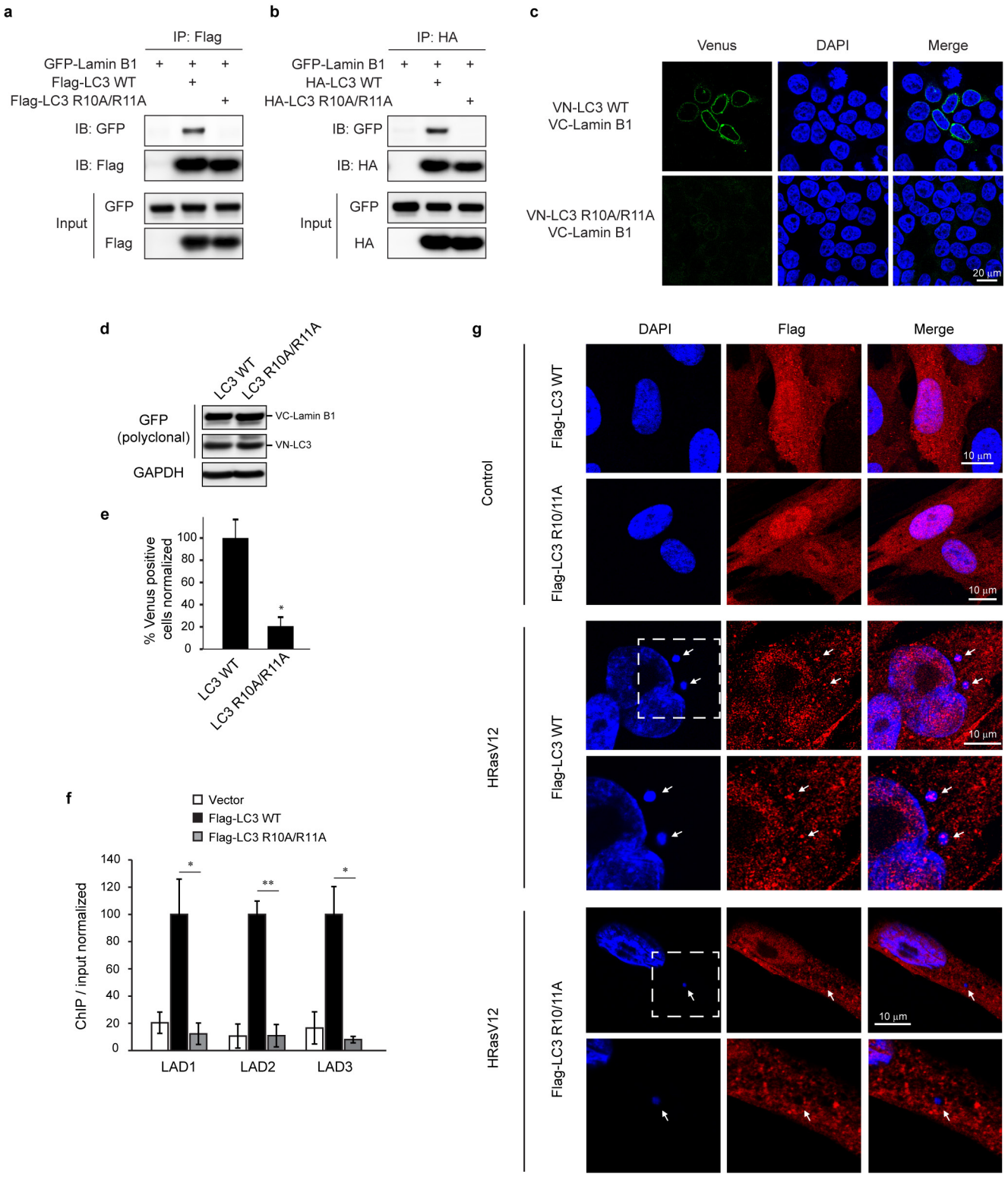
Extended Data Figure 4 | Imaging analyses of mCherry–GFP–lamin B1 HRasV12 cells. a, Related to Fig. 3c. mCherry–GFP–lamin B1 HRasV12 cells stably expressing IMR90 were imaged by three-dimensional super-resolution microscopy. Sections shown span the top, middle, and bottom layers of the cell. The mCherry channel was deliberately under-exposed to prevent over-saturation of the cytoplasmic signals. Scale bar, 5 µm. The insets are presented in Fig. 3c. **b**, Live-cell imaging of mCherry–GFP–lamin B1 HRasV12 IMR90. Images shown are the maximum-projection combining all *z*-sections. Nucleus-to-cytoplasm transport events are labelled sequentially as indicated. Note the initial yellow signal, followed by disappearance of GFP then mCherry, in events 1 and 3; event 2 was not yet degraded by the end of the imaging.



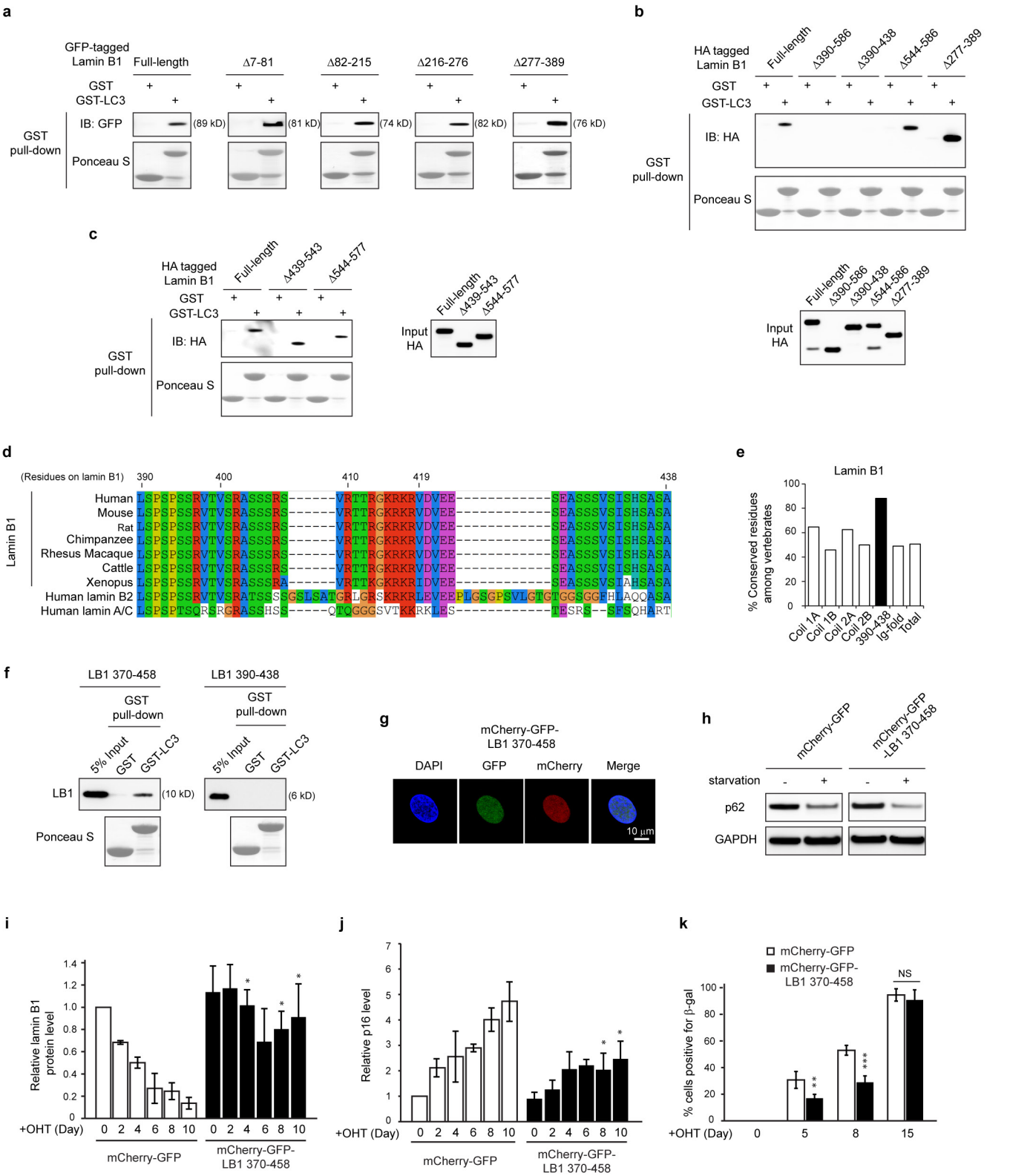
Extended Data Figure 5 | CCF and lamin B1 are targeted by autophagy. **a**, **b**, IMR90 cells stably expressing GFP–LC3 and HRasV12 were stained with indicated antibodies and imaged under confocal microscopy. Cytoplasmic events are labelled by arrows. **c**, HRasV12 IMR90 cells were stained with LC3 antibody. **d**, Related to Fig. 3e, immuno-TEM analysis of GFP–lamin B1 IMR90 cells. Cells were stained with a GFP antibody and conjugated with 10 nm gold particles. Gold particles are indicated by arrows.



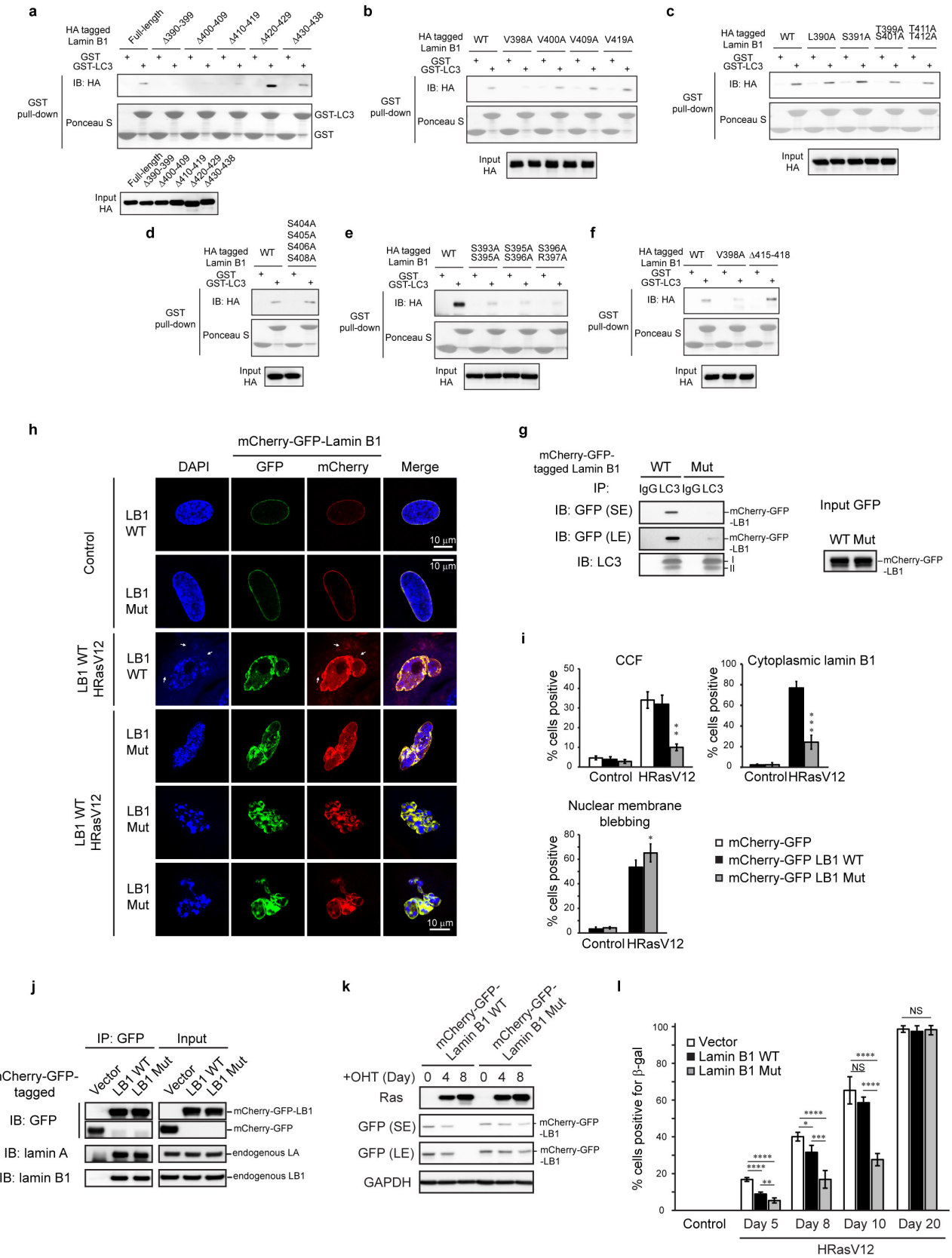
Extended Data Figure 6 | Knockdown of Atg7 attenuates lamin B1 downregulation. a, Related to Fig. 4a, quantification of lamin B1 immunoblots. Bars, mean ± s.e.m.; $n = 3$; $^{*}P < 0.05$, $^{**}P < 0.005$, $^{***}P < 0.0001$, compared with sh-NTC day 0; NS, non-significant. **b**, Reverse transcribed qPCR of cells as in Fig. 4a. Data are the mean normalized to GAPDH ± s.e.m.; $n = 3$. **c**, **d**, IMR90 cells were treated as indicated and analysed by immunoblotting. **e**, BJ cells were treated with etoposide and analysed by immunoblotting. **f**, **g**, Atg7 knockdown inhibits mCherry–GFP–lamin B1 nucleus-to-cytoplasm transport. Bars are mean ± s.d.; $n = 4$, over 100 cells; $^{*}P < 0.0001$. **h**, **i**, ER:HRasV12 BJ cells stably expressing Dox-inducible GFP or GFP–lamin B1 were either left uninduced (bars 1 and 2), or induced with 4-OHT for 3 weeks (3–6). Cells were then induced with Dox (in the presence of 4-OHT) for an additional 2 weeks (5 and 6). **i**, Quantification of β-gal positivity. Bars, mean ± s.d.; $n = 4$, over 200 cells. **j**, Related to Fig. 4a, quantification of p16 immunoblots. Bars, mean ± s.e.m.; $n = 3$; $^{*}P < 0.05$, compared with corresponding sh-NTC controls. **k**, ER:HRas IMR90 cells were scored for β-gal positivity. Bars, mean ± s.d.; $n = 4$, over 200 cells; $^{*}P < 0.0005$, $^{**}P < 0.0001$. One-way ANOVA coupled with Tukey's *post hoc* test for **a** and **i**; all other tests were unpaired two-tailed Student's *t*-tests.



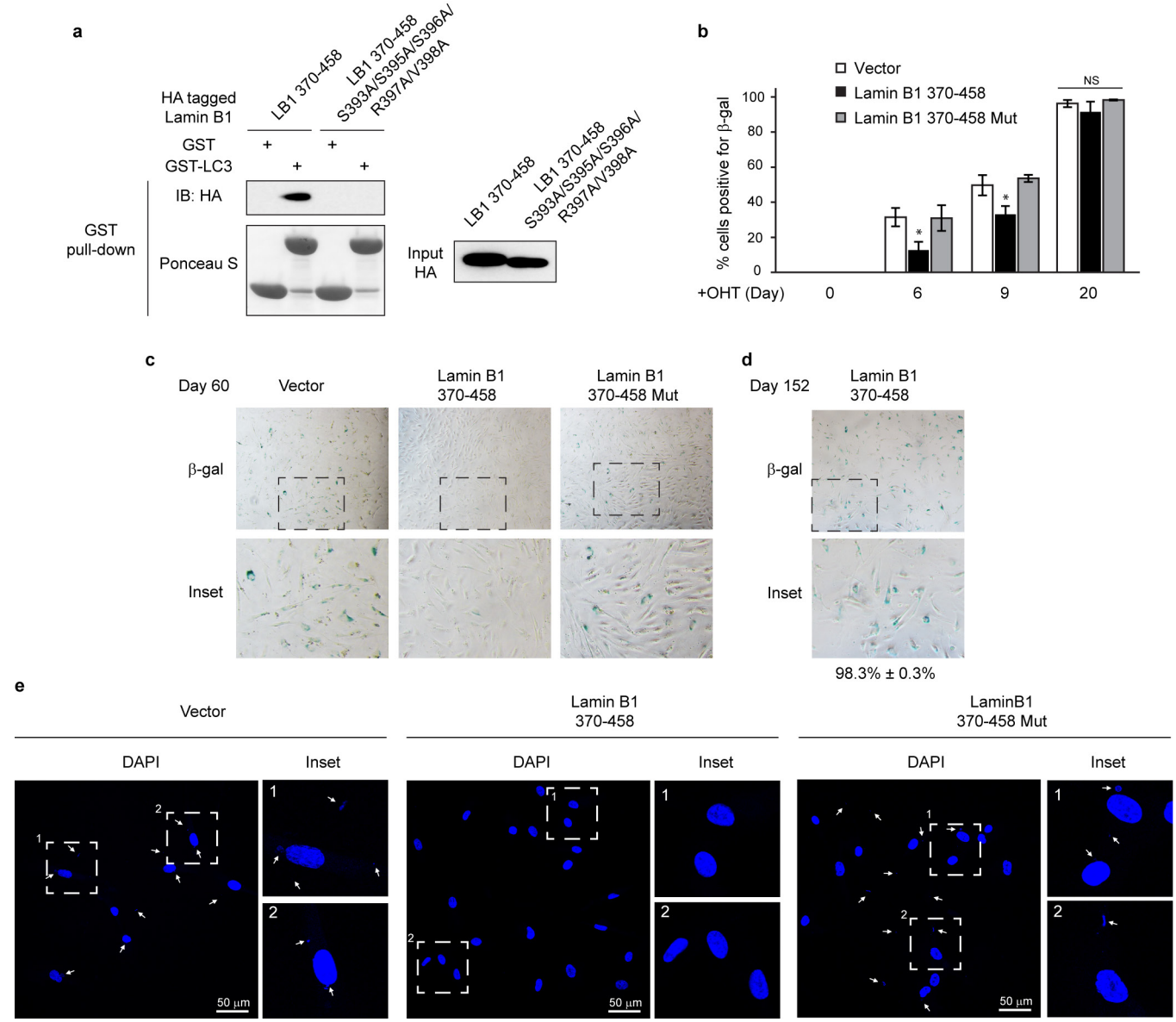
Extended Data Figure 7 | LC3 R10 and R11 are essential for lamin B1 binding. a, **b**, HEK293T cells were transfected as indicated and analysed by co-IP. **c**–**e**, BiFC analyses in HeLa cells transfected with the indicated combination of split Venus constructs. Bars, mean ± s.d.; $n = 4$, over 500 cells; $^{*}P < 0.0001$. **f**, IMR90 cells stably expressing the indicated constructs were analysed by Flag ChIP. Bars, mean ± s.e.m.; $^{*}P < 0.05$, $^{**}P < 0.005$; unpaired two-tailed Student's *t*-test for **e** and **f**. **g**, LC3 R10 and R11 are necessary for co-localization with CCF in HRasV12 IMR90. CCFs are indicated with arrows.



Extended Data Figure 8 | Mapping of LC3–lamin B1 interaction. **a**, HEK293T cells transfected with indicated constructs were analysed by GST–LC3B pull-down. **b**, **c**, *In vitro* translated constructs were subjected to GST–LC3B pull-down. **d**, **e**, Evolutionary analyses of vertebrate lamin B1 and the corresponding regions of other lamin isoforms. **e**, Number of conserved residues normalized to total residues. **f**, Bacterially purified fragments were analysed by GST–LC3B pull-down. **g**, mCherry–GFP–lamin B1 370–458 localizes to the nucleus. **h**, Cells were starved and analysed by immunoblotting. **i**, **j**, Related to Fig. 4f, quantification of lamin B1 and p16 immunoblots; $n = 3$. **k**, ER-HRasV12 IMR90 cells were scored for β-gal positivity; $n = 4$, over 200 cells. Bars, mean ± s.e.m. (**i**, **j**), s.d. (**k**); NS, non-significant; $^{*}P < 0.05$; $^{**}P < 0.0005$; $^{***}P < 0.0001$; unpaired two-tailed Student's *t*-test.



Extended Data Figure 9 | Additional characterization of lamin B1 substitution mutant. a–**f**, Related to Fig. 5a, *in vitro* translated proteins were analysed by GST–LC3B pull-down. **g**, LC3 immunoprecipitation in HEK293T cells transfected as indicated. The remaining interaction with the mutant is probably due to the endogenous lamin B1 that interacts with LC3 and the mutant, as shown in **j**. **h**, **i**, IMR90 cells were imaged under confocal microscopy and quantified. Bars, mean ± s.d.; $n = 4$, over 200 cells; $^{*}P < 0.05$, $^{**}P < 0.005$, $^{***}P < 0.0001$; unpaired two-tailed Student's *t*-test. **j**, HEK293T transfected cells were analysed by immunoprecipitation. **k**, ER:HRasV12 IMR90 cells were induced with OHT and harvested for immunoblotting. **l**, IMR90 cells were quantified for β-gal positivity. Bars, mean ± s.d.; $n = 4$, over 200 cells; $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$, $^{****}P < 0.0001$, NS, non-significant; one-way ANOVA coupled with Tukey's *post hoc* test.



Extended Data Figure 10 | Lamin B1 370–458 fragment extends cellular lifespan. a, *In vitro* translated proteins were analysed by GST–LC3B pull-down. **b**, ER:HRasV12 IMR90 cells were quantified for β-gal positivity. Bars, mean ± s.d.; $n = 4$, over 200 cells; $^*P < 0.05$; NS, non-significant; one-way ANOVA coupled with Tukey's *post hoc* test. **c**, **d**, Related to Fig. 5f, representative images of β-gal. **e**, Related to Fig. 5g, cells were fixed and stained with DAPI. CCFs are indicated by arrows.

LETTER

doi:10.1038/nature15544

Conformational control of DNA target cleavage by CRISPR–Cas9

Samuel H. Sternberg[1], Benjamin LaFrance[2], Matias Kaplan[3]† & Jennifer A. Doudna[1,2,3,4,5]

Cas9 is an RNA-guided DNA endonuclease that targets foreign DNA for destruction as part of a bacterial adaptive immune system mediated by clustered regularly interspaced short palindromic repeats (CRISPR)[1,2]. Together with single-guide RNAs[3], Cas9 also functions as a powerful genome engineering tool in plants and animals[4–6], and efforts are underway to increase the efficiency and specificity of DNA targeting for potential therapeutic applications[7,8]. Studies of off-target effects have shown that DNA binding is far more promiscuous than DNA cleavage[9–11], yet the molecular cues that govern strand scission have not been elucidated. Here we show that the conformational state of the HNH nuclease domain directly controls DNA cleavage activity. Using intramolecular Förster resonance energy transfer experiments to detect relative orientations of the Cas9 catalytic domains when associated with on- and off-target DNA, we find that DNA cleavage efficiencies scale with the extent to which the HNH domain samples an activated conformation. We furthermore uncover a surprising mode of allosteric communication that ensures concerted firing of both Cas9 nuclease domains. Our results highlight a proofreading mechanism beyond initial protospacer adjacent motif (PAM) recognition[12] and RNA–DNA base-pairing[3] that serves as a final specificity checkpoint before DNA double-strand break formation.

Cas9 is a large, multi-domain protein that undergoes RNA-induced conformational changes to reach a DNA-binding-competent state[13]. Crystal structures of *apo*[13], single-guide RNA (sgRNA)-bound[14], and sgRNA/DNA-bound[15,16] Cas9 from *Streptococcus pyogenes* (Fig. 1a, b) have revealed distinct conformational states of the protein but failed to explain its DNA cleavage mechanism, because in each structure the HNH domain active site is positioned at least 30 Å away from the DNA cleavage site[15,16]. Furthermore, available structures could not explain why DNA cleavage is precluded at stably bound off-target sites with incomplete RNA-DNA complementarity. We hypothesized that functionally important HNH conformational dynamics could influence the cleavage specificity of the Cas9–guide RNA enzyme complex. To test this possibility, we developed a Förster resonance energy transfer (FRET)-based approach to investigate Cas9 structural changes in response to binding sgRNA and DNA ligands.

We generated a FRET construct to monitor Cas9 structural rearrangements upon sgRNA binding[13] (Fig. 1b). Starting with a cysteine-free Cas9 variant, we introduced cysteine residues at positions D435 and E945 near the hinge region and labelled these residues with Cy3- and Cy5-maleimide dyes, generating $Cas9_{hinge}$. Control labelling reactions with cysteine-free Cas9 confirmed the conjugation specificity, and doubly labelled Cas9 was fully functional for DNA cleavage (Extended Data Fig. 1a–c). Measurements from available structures revealed an expected distance change of ~60 Å upon sgRNA and DNA binding (Extended Data Table 1), and indeed, when Cy3 of sgRNA-bound $Cas9_{hinge}$ was excited at 530 nm, we observed a substantial decrease in energy transfer compared with *apo*-$Cas9_{hinge}$, as evidenced by a relative increase in donor (Cy3) fluorescence relative to acceptor (Cy5) fluorescence (Fig. 1c). The observed change scaled with the molar ratio of sgRNA to Cas9, a mixture of donor-only and acceptor-only labelled $Cas9_{hinge}$ showed no evidence of energy transfer, and an sgRNA specific to *Neisseria meningitidis* Cas9 (ref. 17), which significantly impairs *S. pyogenes* Cas9 binding (data not shown), elicited a negligible change (Extended Data Fig. 2a–c). We conclude that the change in fluorescence intensities resulted from an sgRNA-induced, intramolecular conformational change in $Cas9_{hinge}$.

$Cas9_{hinge}$ exhibited an ~70% decrease in energy transfer upon sgRNA binding as determined by $(ratio)_A$, whereby the acceptor fluorescence intensity via energy transfer is normalized to that via direct excitation[18,19] (Methods and Extended Data Fig. 2d). Target DNA binding induced little further change in FRET (Fig. 1c, d), consistent with available structural data (Extended Data Table 1). To identify the molecular

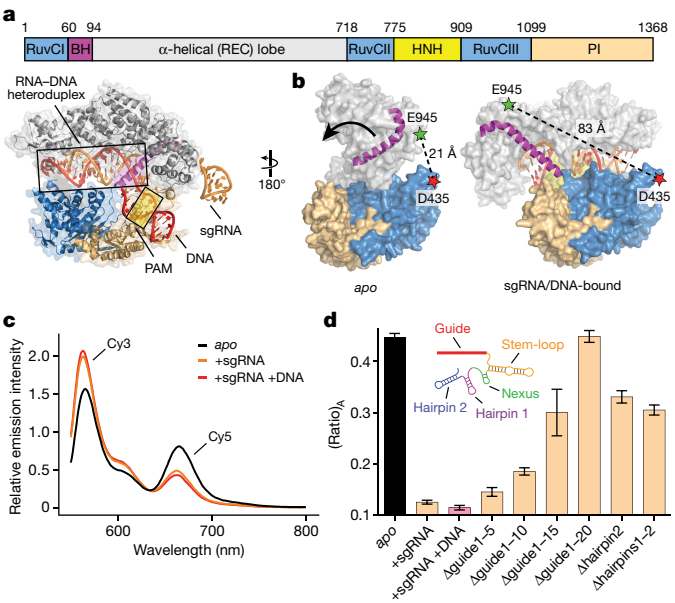


Figure 1 | Full-length sgRNA drives inward lobe closure of Cas9. **a**, Domain organization of *S. pyogenes* Cas9 (top) and X-ray crystal structure of sgRNA/DNA-bound Cas9 (Protein Data Bank (PDB) accession number 4UN3, ref. 16) (bottom), with HNH domain omitted for clarity. BH, bridge helix; REC, recognition; PI, PAM-interacting. **b**, Design of $Cas9_{hinge}$ FRET construct. Measured distances between D435 and E945 in *apo* (PDB 4CMP, ref. 13) and sgRNA/DNA-bound Cas9 structures are indicated. Inward lobe closure is exemplified by movement of the BH (arrow). Regions of the PI domain, sgRNA, and DNA are omitted for clarity. **c**, Fluorescence emission spectra for $Cas9_{hinge}$ in the presence of the indicated substrates. **d**, $(Ratio)_A$ data for $Cas9_{hinge}$. Inset: schematic of full-length sgRNA coloured by motif[28]. Error bars, s.d.; $n = 3$.

[1]Department of Chemistry, University of California, Berkeley, California 94720, USA. [2]Department of Molecular and Cell Biology, University of California, Berkeley, California 94720, USA. [3]Howard Hughes Medical Institute, University of California, Berkeley, California 94720, USA. [4]Innovative Genomics Initiative, University of California, Berkeley, California 94720, USA. [5]Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA. †Present address: Department of Bioengineering, Stanford University, Stanford, California 94305, USA.

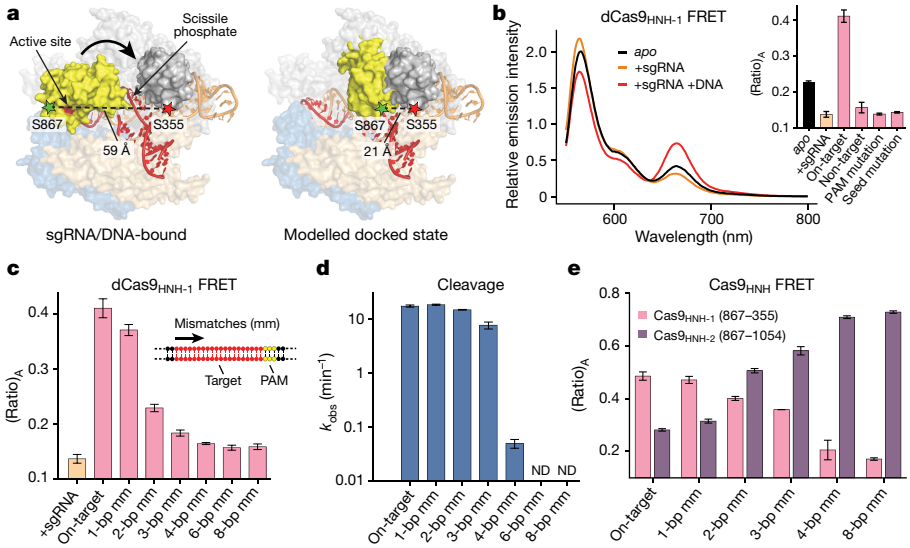


Figure 2 | FRET experiments reveal an activated conformation of the HNH nuclease domain. a, Design of $Cas9_{HNH-1}$ FRET construct. Measured distances between S355 and S867 in the sgRNA/DNA-bound Cas9 structure[16] and a model of the HNH domain docked at the cleavage site are indicated, as are putative conformational changes of the HNH domain (arrow). The model was generated using an HNH homologue structure (PDB 2QNC, ref. 21). **b**, Fluorescence emission spectra for $dCas9_{HNH-1}$ in the presence of the indicated substrates. Inset: $(ratio)_A$ values. **c**, $(Ratio)_A$ data for $dCas9_{HNH-1}$. Mismatches (mm) were introduced sequentially from the PAM-distal end of the target. **d**, Cleavage rate constants using wild-type Cas9. ND, cleavage not detected. **e**, $(Ratio)_A$ data for catalytically active $Cas9_{HNH-1}$ and $Cas9_{HNH-2}$. Error bars in **b**–**e**, s.d.; $n = 3$.

determinants that trigger conformational rearrangement of Cas9, we tested truncated variants of the sgRNA (Extended Data Table 2) and found that the 20-nucleotide target recognition sequence has a critical role in controlling the Cas9 conformational state (Fig. 1d). An sgRNA lacking the entire guide segment (Δguide1–20) generated a $(ratio)_A$ value indistinguishable from *apo*-$Cas9_{hinge}$ while being more than 95% bound under our experimental conditions[14], whereas sgRNAs containing part of the 20-nucleotide guide segment partly restored the change in $(ratio)_A$. sgRNA variants lacking one or both hairpins at the 3′ end (Δhairpins1–2) also generated intermediate $(ratio)_A$ values (Fig. 1d) while retaining sub-nanomolar binding affinity to Cas9 (ref. 20), and similar data were obtained with catalytically dead (D10A/H840A) $dCas9_{hinge}$ (Extended Data Fig. 2e). We conclude that motifs at both ends of the sgRNA are required to stabilize a closed state of Cas9, but that in the case of Δhairpins1–2, a fully closed state is not required for rapid cleavage kinetics[20]. We propose that intermediate $(ratio)_A$ changes reflect stable sgRNA–$Cas9_{hinge}$ complexes interconverting between open and closed conformers.

We next focused on the HNH nuclease domain. Since existing crystal structures exhibit inactive HNH domain conformations[15,16], we built a model for the putative activated state by docking a homologous HNH–dsDNA crystal structure[21] onto the sgRNA/DNA-bound Cas9 structure (Extended Data Fig. 3a–d). We selected two pairs of positions (S355–S867 and S867–N1054) whose inter-residue distances, according to our model, would change substantially upon target DNA binding (Fig. 2a, Extended Data Fig. 3e and Extended Data Table 1). Cas9 labelled with Cy3 and Cy5 at these sites ($Cas9_{HNH-1}$ and $Cas9_{HNH-2}$) retained nearly wild-type DNA cleavage activity (Extended Data Fig. 1c).

We observed a substantial FRET increase for catalytically inactive $dCas9_{HNH-1}$ upon target DNA binding relative to sgRNA alone (Fig. 2b), and control experiments with non-target DNA or off-target DNA substrates containing either PAM or seed mutations failed to generate this change (Fig. 2b and Extended Data Table 2). We next monitored FRET with off-target DNA substrates containing mutations distal from the PAM, which retain high-affinity Cas9 binding[12,22]. Remarkably, the observed $(ratio)_A$ values decreased as the number of mismatches increased (Fig. 2c), and these changes were not attributable to decreasing occupancy of the sgRNA/DNA-bound complex: direct binding assays indicate at least 89% of the $dCas9_{HNH-1}$ population should be bound to all tested DNA substrates, and increasing the concentration of dsDNA had no discernible effect on $(ratio)_A$ (Extended Data Fig. 4a, b). Our results show that the HNH domain samples a conformational equilibrium with on-target DNA that is distinct from partly matching off-target DNA, and suggest that the high FRET state corresponds to an active HNH conformation at the cleavage site.

We suspected that altered conformational states of the HNH domain could explain which off-target DNA substrates are cleaved by Cas9. Substrates with at least 4 base-pair (bp) mismatches that elicited a low $(ratio)_A$ value were cleaved slowly, if at all (Fig. 2d and Extended Data Fig. 4c), as observed previously[22,23]. This indicates that the inability to access the high FRET state associated with an activated HNH conformation precludes cleavage. Interestingly, substrates with only 1–3 bp mismatches at the distal end of the target sequence were cleaved at near wild-type rates despite having diminished $(ratio)_A$ values relative to the on-target. This suggests that rapidly interconverting conformational states, one of which is the activated state, may still enable rapid cleavage. Truncated sgRNAs with shorter regions of target complementarity that exhibit enhanced fidelity in genome editing experiments[24] may similarly facilitate efficient on-target cleavage without stabilizing an activated HNH conformation. Single-molecule experiments will be necessary to reveal these putative dynamics, which are unavoidably averaged in our ensemble measurements.

We observed a similar pattern of $(ratio)_A$ changes using catalytically active $Cas9_{HNH}$, and the opposite trend of $(ratio)_A$ changes was observed with $Cas9_{HNH-2}$, a construct designed to undergo a high-to-low FRET efficiency transition upon on-target DNA binding (Fig. 2e and Extended Data Figs 3e and 4d). These data suggest that positioning of the HNH domain is largely unaffected by actual strand scission, but instead reflects a conformational equilibrium that is particularly sensitive to RNA–DNA heteroduplex formation at the distal end of the target. These observations emphasize the importance of RNA–DNA complementarity throughout the target region, rather than only the seed sequence closest to the PAM, in controlling Cas9 cleavage specificity.

The HNH and RuvC nuclease domains cleave target and non-target strands 3 bp upstream of the PAM, respectively[3,25]. For partly unwound off-target substrates with mismatches >10 bp further upstream, target strand cleavage is precluded by conformational control of the HNH domain. However, the mechanism by which RuvC domain-catalysed non-target strand cleavage is avoided remains unknown.

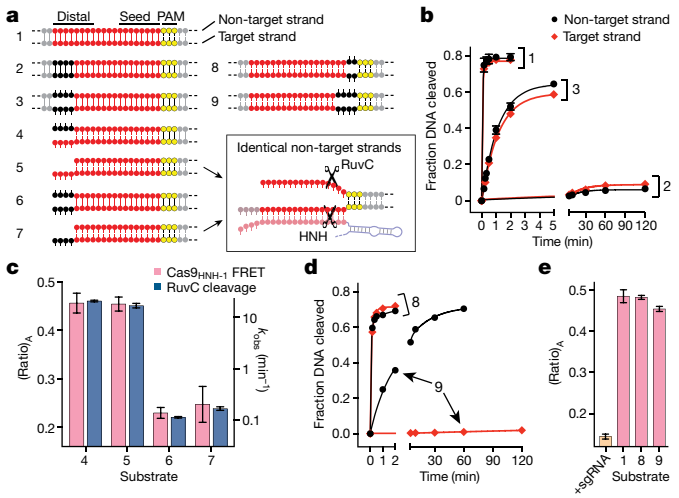


Figure 3 | RuvC nuclease activity is allosterically controlled by HNH conformational changes. a, Tested DNA substrates, with on-target (1) at top. Matched and mismatched positions of DNA target strand sequences relative to the sgRNA are coloured red and black, respectively, with the PAM in yellow. Some substrates contain internal mismatches between the two DNA strands; dashed lines indicate additional flanking sequences. Schematic at bottom right depicts identical non-target strand substrates presented to the RuvC nuclease domain in substrates 5 and 7. **b**, Non-target (black) and target (red) strand cleavage time courses for the indicated DNA substrates using wild-type Cas9. Exponential fits are shown as solid lines. **c**, $(Ratio)_A$ data for $Cas9_{HNH-1}$ (pink bars, left *y* axis) and non-target strand cleavage kinetics of the RuvC domain (blue bars, right *y* axis) for the indicated DNA substrates. **d**, Non-target and target strand cleavage time courses for the indicated DNA substrates using wild-type Cas9. Exponential fits are shown as solid lines. **e**, $(Ratio)_A$ data for $Cas9_{HNH-1}$. Error bars in **b**–**e**, s.d.; $n = 3$.

We hypothesized that this activity would be sensitive to HNH domain conformational changes. We first separately measured HNH and RuvC domain cleavage rates for a panel of partly mismatched substrates and found that both strands were consistently cleaved in synchrony (Fig. 3a, b and Extended Data Fig. 5a, b). We next used shorter DNA substrates with or without internal mismatches, such that Cas9-mediated DNA unwinding up to the site of an sgRNA–DNA mismatch would theoretically present identical substrates to the RuvC domain active site (Fig. 3a). After separately measuring non-target strand cleavage kinetics and $Cas9_{HNH-1}$ FRET, we observed a tight correlation between RuvC domain cleavage activity and the presence of an activated HNH conformational state (Fig. 3c and Extended Data Fig. 5c–e). This finding provides strong evidence that HNH conformational dynamics exert allosteric control over the RuvC nuclease domain. Furthermore, the RuvC domain could still effectively cleave the non-target strand of a bubbled substrate that induced an activated HNH conformation, but whose target strand could not be cleaved by the HNH domain because of mismatches in the seed (Fig. 3d, e). Together, these data argue that HNH conformational changes, but not HNH nuclease function, trigger RuvC domain nuclease activity.

We wondered how Cas9 achieves this functional coupling. The HNH domain is inserted between RuvC domain motifs II and III, but linkers connecting both domains are consistently disordered in available crystal structures and there are relatively few inter-domain contacts[13,15,16] (Extended Data Fig. 6a). We purified an HNH deletion construct, ΔHNH–Cas9 (Extended Data Fig. 6a–c), that retained nearly wild-type DNA binding activity while being defective in non-target strand cleavage by the RuvC domain (Fig. 4a, b and Extended Data Fig. 6d). Thus, the HNH domain is required for RuvC nuclease domain activation but is dispensable for RNA-guided DNA targeting.

Finally, we sought to identify the basis of allostery between the HNH and RuvC domains. We hypothesized that two α-helices connecting the HNH and RuvC III motifs (residues S909–N940), previously shown to adopt an extended conformation and proposed to assist the HNH domain in approaching the cleavage site[15], were instead acting as a signal transducer (Extended Data Fig. 7a). We introduced a series of proline residues to specifically disrupt this α-helix and found that target strand cleavage kinetics by the HNH domain were minimally affected (Fig. 4c–e and Extended Data Fig. 7b, c). In stark contrast, RuvC domain nuclease activity was almost completely blocked with an E923P/T924P–Cas9 mutant, and this effect could be reversed with the corresponding alanine mutations (Fig. 4d, e and Extended Data Fig. 7c). The finding that this effect was not confined to highly conserved residues supports the idea that disruption of the helix-forming propensity of this region, and not specific point mutations, disabled the RuvC domain. We conclude that an intact extended α-helix acts as an allosteric switch to communicate the HNH conformational change to the RuvC domain and activate it for cleavage. Understanding the precise mechanism of activation will probably require additional structures of Cas9 in a pre-cleavage state, with the intact non-target strand substrate bound in the RuvC active site.

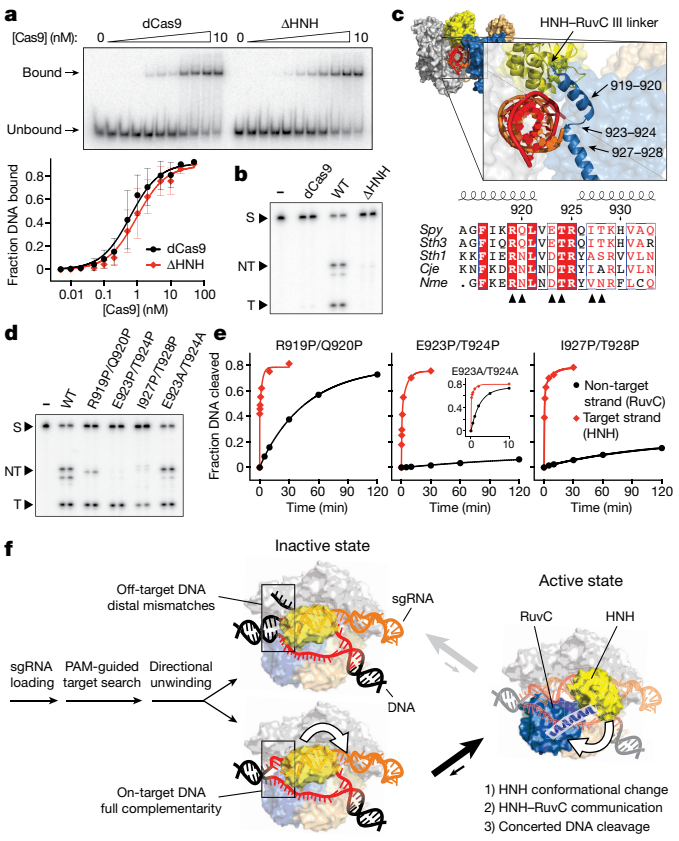


Figure 4 | Mechanism of communication between the HNH and RuvC nuclease domains to achieve concerted DNA cleavage. **a**, Target DNA binding assay with dCas9 and ΔHNH–Cas9, resolved by native polyacrylamide gel electrophoresis (PAGE) (top); for gel source data, see Supplementary Fig. 1. Quantified data are below; binding fits are shown as solid lines. **b**, Target DNA cleavage assay with dCas9, wild-type (WT) Cas9, and ΔHNH–Cas9, resolved by denaturing PAGE. S, substrate; NT, cleaved non-target strand; T, cleaved target strand. **c**, Magnified view of the sgRNA/DNA-bound Cas9 structure[16] (top) highlights two α-helices connecting the HNH domain carboxy (C) terminus and RuvC III amino (N) terminus. Bottom shows sequence alignment[29] of this region, and residues mutated to proline or alanine are indicated (arrows). **d**, Target DNA cleavage assay with the indicated Cas9 variants, resolved by denaturing PAGE. **e**, Target (red) and non-target (black) strand cleavage time courses with the indicated Cas9 variants (for WT-Cas9 data, see Fig. 3b). Exponential fits are shown as solid lines. Error bars in **a** and **e**, s.d.; $n = 5$ and 3, respectively. **f**, Model for conformational control of target cleavage by CRISPR–Cas9.

Our data support a model in which the Cas9 endonuclease uses multiple levels of regulation to ensure accurate target DNA cleavage (Fig. 4f and Supplementary Video 1). After identification of potential targets via PAM binding and directional DNA unwinding dependent on sgRNA–DNA complementarity, recognition of on-target DNA drives a conformational change in the HNH nuclease domain that enables productive engagement with the scissile phosphate. Importantly, this same structural transition triggers RuvC domain catalytic activity, ensuring concerted cleavage of both DNA strands. Partly complementary off-target DNA sequences may stably bind Cas9, but by failing to drive HNH conformational changes, avoid cleavage. The recent crystal structure of sgRNA/DNA-bound Cas9 from *Staphylococcus aureus* (~17% sequence identity with *S. pyogenes* Cas9) also exhibits an inactive HNH conformation[26], suggesting that conformational control of the HNH domain is a general feature of all Cas9 enzymes. Furthermore, this proofreading mechanism is strikingly similar to the R-loop locking mechanism used by the RNA-guided targeting complex (Cascade) from type I CRISPR-Cas systems, in which RNA–DNA heteroduplex formation at the PAM-distal end of the target exerts allosteric control over Cascade conformational rearrangements near the PAM-proximal end that are required for subsequent target cleavage[22,27]. Beyond providing fundamental insights into the mechanism of DNA interrogation by Cas9, our findings have important implications for the use of Cas9 as a genome engineering technology. For example, our data can explain why little cleavage occurs at off-target DNA sequences identified in chromatin immunoprecipitation followed by sequencing (ChIP-seq) experiments[9–11], and suggest that DNA nicking by the native Cas9 enzyme is disfavoured in cells owing to concerted cutting by the HNH and RuvC nuclease domains. Finally, our findings demonstrate an exciting opportunity to use protein conformational changes that report on target DNA recognition for fluorescence-based readout of DNA binding in cells.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 28 June; accepted 1 September 2015.
Published online 28 October 2015.

1. van der Oost, J., Westra, E. R., Jackson, R. N. & Wiedenheft, B. Unravelling the structural and mechanistic basis of CRISPR-Cas systems. *Nature Rev. Microbiol.* **12,** 479–492 (2014).
2. Barrangou, R. & Marraffini, L. A. CRISPR-Cas systems: prokaryotes upgrade to adaptive immunity. *Mol. Cell* **54,** 234–244 (2014).
3. Jinek, M. *et al.* A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **337,** 816–821 (2012).
4. Hsu, P. D., Lander, E. S. & Zhang, F. Development and applications of CRISPR-Cas9 for genome engineering. *Cell* **157,** 1262–1278 (2014).
5. Doudna, J. A. & Charpentier, E. Genome editing. The new frontier of genome engineering with CRISPR-Cas9. *Science* **346,** 1258096 (2014).
6. Sternberg, S. H. & Doudna, J. A. Expanding the biologist's toolkit with CRISPR-Cas9. *Mol. Cell* **58,** 568–574 (2015).
7. Wu, X., Kriz, A. J. & Sharp, P. A. Target specificity of the CRISPR-Cas9 system. *Quant. Biol.* **2,** 59–70 (2014).
8. Gori, J. L. *et al.* Delivery and specificity of CRISPR-Cas9 genome editing technologies for human gene therapy. *Hum. Gene Ther.* **26,** 443–451 (2015).
9. Wu, X. *et al.* Genome-wide binding of the CRISPR endonuclease Cas9 in mammalian cells. *Nature Biotechnol.* **32,** 670–676 (2014).
10. Tsai, S. Q. *et al.* GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR-Cas nucleases. *Nature Biotechnol.* **33,** 187–197 (2015).
11. Ran, F. A. *et al. In vivo* genome editing using *Staphylococcus aureus* Cas9. *Nature* **520,** 186–191 (2015).
12. Sternberg, S. H., Redding, S., Jinek, M., Greene, E. C. & Doudna, J. A. DNA interrogation by the CRISPR RNA-guided endonuclease Cas9. *Nature* **507,** 62–67 (2014).
13. Jinek, M. *et al.* Structures of Cas9 endonucleases reveal RNA-mediated conformational activation. *Science* **343,** 1247997 (2014).
14. Jiang, F., Zhou, K., Ma, L., Gressel, S. & Doudna, J. A. A Cas9–guide RNA complex preorganized for target DNA recognition. *Science* **348,** 1477–1481 (2015).
15. Nishimasu, H. *et al.* Crystal structure of Cas9 in complex with guide RNA and target DNA. *Cell* **156,** 935–949 (2014).
16. Anders, C., Niewoehner, O., Duerst, A. & Jinek, M. Structural basis of PAM-dependent target DNA recognition by the Cas9 endonuclease. *Nature* **513,** 569–573 (2014).
17. Hou, Z. *et al.* Efficient genome engineering in human pluripotent stem cells using Cas9 from *Neisseria meningitidis*. *Proc. Natl Acad. Sci. USA* **110,** 15644–15649 (2013).
18. Majumdar, Z. K., Hickerson, R., Noller, H. F. & Clegg, R. M. Measurements of internal distance changes of the 30S ribosome using FRET with multiple donor-acceptor pairs: quantitative spectroscopic methods. *J. Mol. Biol.* **351,** 1123–1145 (2005).
19. Clegg, R. M. Fluorescence resonance energy transfer and nucleic acids. *Methods Enzymol.* **211,** 353–388 (1992).
20. Wright, A. V. *et al.* Rational design of a split-Cas9 enzyme complex. *Proc. Natl Acad. Sci. USA* **112,** 2984–2989 (2015).
21. Biertümpfel, C., Yang, W. & Suck, D. Crystal structure of T4 endonuclease VII resolving a Holliday junction. *Nature* **449,** 616–620 (2007).
22. Szczelkun, M. D. *et al.* Direct observation of R-loop formation by single RNA-guided Cas9 and Cascade effector complexes. *Proc. Natl Acad. Sci. USA* **111,** 9798–9803 (2014).
23. Cencic, R. *et al.* Protospacer adjacent motif (PAM)-distal sequences engage CRISPR Cas9 DNA target cleavage. *PLoS One* **9,** e109213 (2014).
24. Fu, Y., Sander, J. D., Reyon, D., Cascio, V. M. & Joung, J. K. Improving CRISPR-Cas nuclease specificity using truncated guide RNAs. *Nature Biotechnol.* **32,** 279–284 (2014).
25. Gasiunas, G., Barrangou, R., Horvath, P. & Siksnys, V. Cas9-crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria. *Proc. Natl Acad. Sci. USA* **109,** E2579–E2586 (2012).
26. Nishimasu, H. *et al.* Crystal structure of *Staphylococcus aureus* Cas9. *Cell* **162,** 1113–1126 (2015).
27. Rutkauskas, M. *et al.* Directional R-loop formation by the CRISPR-Cas surveillance complex cascade provides efficient off-target site rejection. *Cell Reports* **10,** 1534–1543 (2015).
28. Briner, A. E. *et al.* Guide RNA functional modules direct Cas9 activity and orthogonality. *Mol. Cell* **56,** 333–339 (2014).
29. Robert, X. & Gouet, P. Deciphering key features in protein structures with the new ENDscript server. *Nucleic Acids Res.* **42,** W320–W324 (2014).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank D. Taylor and J. Chen for discussions, M. O'Connell, L. Ma, N. Ma, and K. Zhou for technical assistance, and members of the Doudna laboratory for reading the manuscript. S.H.S. acknowledges support from the National Science Foundation and National Defense Science & Engineering Graduate Research Fellowship programs. B.L. acknowledges support from a National Institutes of Health National Research Service Award Training Grant (T32GM007232). J.A.D. is an Investigator of the Howard Hughes Medical Institute.

Author Contributions S.H.S. designed and conducted all experiments. B.L. and M.K. assisted with protein purification, dye labelling, cleavage assays, and fluorescence experiments. All authors discussed the data; S.H.S. and J.A.D. wrote the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details are available in the online version of the paper. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to J.A.D. (doudna@berkeley.edu).

METHODS

No statistical methods were used to predetermine sample size. The experiments were not randomized. The investigators were not blinded to allocation during experiments and outcome assessment.

Cas9 and nucleic acid preparation. *S. pyogenes* Cas9 was cloned into a custom pET-based expression vector encoding an N-terminal His_{10}-tag followed by maltose-binding protein (MBP) and a TEV protease cleavage site. Point mutations were introduced using site-directed mutagenesis or around-the-horn PCR and verified by DNA sequencing. dCas9 refers to catalytically inactive (dead) Cas9 containing D10A and H840A mutations. ΔHNH-Cas9 contained a deletion of residues T769–K918 and replacement with a GGSGGS linker. The HNH domain for add-back experiments (Extended Data Fig. 6d) encoded residues N776–G907. All Cas9 variants were purified as described[13].

sgRNA templates were PCR amplified and cloned into EcoRI and BamHI sites in pUC19, and encoded full-length CRISPR RNA (crRNA) and trans-activating crRNA (tracrRNA) sequences connected via a GAAA tetraloop (Extended Data Table 2). sgRNAs were transcribed *in vitro* as described[30] and purified using 5–10% denaturing PAGE.

DNA substrates (Extended Data Table 2) were prepared from commercially synthesized oligonucleotides (Integrated DNA Technologies). DNA duplexes without internal mismatches were prepared and purified by native PAGE as described[13]. DNA duplexes containing internal mismatches or overhangs were prepared by mixing a 5× molar excess of one strand with its complementary strand in hybridization buffer (20 mM Tris-Cl pH 7.5, 100 mM KCl, 5 mM $MgCl_2$), heating at 95 °C for 1–2 min, and slow-cooling on the benchtop. For FRET experiments, the non-target strand was in excess over the target strand; for biochemical cleavage experiments, the non-radiolabelled strand was in excess over the radiolabelled strand.

Preparation of dye-labelled Cas9. Labelling reactions were conducted in Cas9 gel filtration buffer (20 mM Tris-Cl pH 7.5, 200 mM KCl, 5% glycerol, 1 mM TCEP) and contained 10 μM Cas9 and 200 μM Cy3- and Cy5-maleimide (GE Healthcare). Dyes were initially dissolved in anhydrous DMSO before being mixed with Cas9, and the final DMSO concentration did not exceed 5%. Reactions were incubated in the dark for 2 h at room temperature (~22 °C) followed by incubation overnight at 4 °C. Reactions were quenched by adding 10 mM DTT, and labelled Cas9 was separated from free dye by size-exclusion chromatography on a Superdex 200 10/300 column. Samples were then concentrated, snap frozen in liquid nitrogen, and stored at −80 °C. Control labelling reactions contained either cysteine-free Cas9 or only one of the two dyes.

FRET experiments. All fluorescence measurements were conducted at room temperature in reaction buffer (20 mM Tris-Cl pH 7.5, 100 mM KCl, 5 mM $MgCl_2$, 5% glycerol, 1 mM DTT), supplemented with 50 μg ml^{-1} heparin to reduce non-specific DNA binding[12]. Reactions (60 μl) with $Cas9_{hinge}$ (C80S/D435C/C574S/E945C-Cas9 labelled with Cy3/Cy5) and $dCas9_{hinge}$ ($Cas9_{hinge}$ with additional nuclease-inactivating D10A/H840A mutations) contained either 50 nM or 100 nM Cas9, and, when present, a 10× and 4× molar excess of sgRNA and target DNA, respectively. Reactions (60 μl) with $Cas9_{HNH\text{-}1}$ (C80S/S355C/C574S/S867C-Cas9 labelled with Cy3/Cy5), $dCas9_{HNH\text{-}1}$ ($Cas9_{HNH\text{-}1}$ with additional nuclease-inactivating D10A/H840A mutations), and $Cas9_{HNH\text{-}2}$ (C80S/C574S/S867C/N1054C-Cas9 labelled with Cy3/Cy5) contained 50 nM Cas9, and, when present, 200 nM sgRNA and DNA unless otherwise indicated.

We observed substantial aggregation of *apo*-Cas9 upon 10 min incubation at 37 °C, as indicated by apparent intermolecular FRET with a single-cysteine Cas9 (C80S/C574S/S867C) that had been labelled with a mixture of Cy3- and Cy5-maleimide (data not shown). This aggregation could be completely avoided by incubating reactions for 10 min at room temperature instead, centrifuging reactions for 5 min at 16,000*g* and 4 °C, and using the supernatant for subsequent fluorescence measurements. This binding protocol was used for all reported FRET data. Reactions were kept at room temperature for about 10–100 min before acquisition of fluorescence spectra, and this variable time delay had no effect on the resulting data, even for reactions with catalytically active Cas9 (data not shown).

Fluorescence measurements were collected with a 3 mm path-length quartz cuvette (Hellma Analytics) and a FluoroMax-3 (HORIBA Jobin Yvon), using 5 nm slit widths and 0.2 s integration time. For each sample, two fluorescence emission spectra were recorded: (1) the sample was excited at 530 nm and emitted light was collected from 550–800 nm in 1 nm increments; and (2) the sample was excited at 630 nm and emitted light was collected from 650–800 nm in 1 nm increments. Data processing was conducted using FluorEssence software (HORIBA Jobin Yvon). Experiments were replicated at least three times, and the presented data are representative results unless stated otherwise.

FRET analysis. The distance between donor and acceptor dyes can be directly calculated from the FRET efficiency, but accurately relating these variables requires knowledge of numerous complex parameters[18]. Our labelling strategy resulted in a heterogeneous mixture of unlabelled, singly-, and doubly-labelled species, further complicating the analysis. We therefore report $(ratio)_A$ as defined by refs 18, 19, whereby acceptor (Cy5) fluorescence via energy transfer is normalized against acceptor fluorescence via direct excitation (Extended Data Fig. 2d), without pursuing a more rigorous calculation of exact distances. $(Ratio)_A$ is directly proportional to FRET efficiency, and changes in $(ratio)_A$ across different experimental conditions serve as a proxy for conformational changes.

For each FRET construct, a donor-only (Cy3-labelled) sample was prepared and its emission spectrum in the *apo* state after 530 nm excitation collected. This spectrum was normalized to and subtracted from each experimental emission spectrum to generate an extracted fluorescence spectrum for the acceptor via energy transfer. The integrated area under this curve from 650–800 nm was calculated and divided by the integrated area under the curve of a spectrum resulting from direct acceptor excitation at 630 nm, resulting in $(ratio)_A$, the enhancement of acceptor fluorescence due to FRET. Raw fluorescence emission spectra presented in the figures and Extended Data figures were normalized and smoothed using the Savistsky–Golay method, and all data analysis used Prism (GraphPad Software).

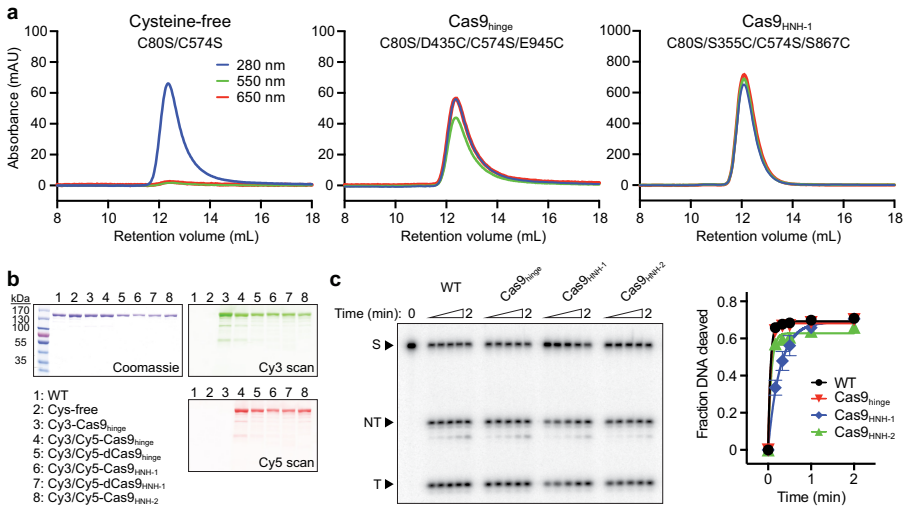
DNA binding and cleavage assays. Biochemical assays were conducted essentially as described[13]. Binding reactions used <0.1 nM 5′-[^{32}P]DNA duplex substrates radiolabelled on both strands and a constant excess of 100 nM sgRNA in the presence of increasing concentrations of dCas9 or ΔHNH-Cas9. Cas9 and sgRNA were pre-incubated at 37 °C for 10 min in reaction buffer supplemented with 50 μg ml^{-1} heparin before being incubated with DNA for ~30 min at room temperature. Reactions were resolved by 5% native PAGE (0.5× TBE, 5 mM $MgCl_2$) at 4 °C and visualized by phosphorimaging (GE Healthcare).

DNA cleavage experiments presented in Figs 2d and 4b, d and Extended Data Figs 1c and 6d used 5′-[^{32}P]DNA duplex substrates radiolabelled on both strands; all other cleavage experiments used DNA duplex substrates with a single 5′-[^{32}P]-radiolabelled strand that had been annealed to a 5× molar excess of unlabelled complementary strand. Cas9 and sgRNA were pre-incubated at 37 °C for 10 min in reaction buffer before adding DNA. Cleavage reactions were performed at room temperature and contained 1 nM DNA and 100 nM Cas9–sgRNA complex. Aliquots were removed at various time points and quenched by mixing with an equal volume of formamide gel loading buffer supplemented with 50 mM EDTA. Cleavage products were resolved by 10% denaturing PAGE and visualized by phosphorimaging (GE Healthcare). Reported pseudo-first-order rate constants (k_{obs}) represent the population-weighted average from double-exponential fits (target strand cleavage data for proline mutants, Fig. 4e and Extended Data Fig. 7b, c) or the result from single-exponential fits (all other data). In some cases, where the observed fraction of cleaved DNA was <0.1 after 2 h, the exponential fit plateau was fixed at 0.75 to avoid overestimating the rate constant.

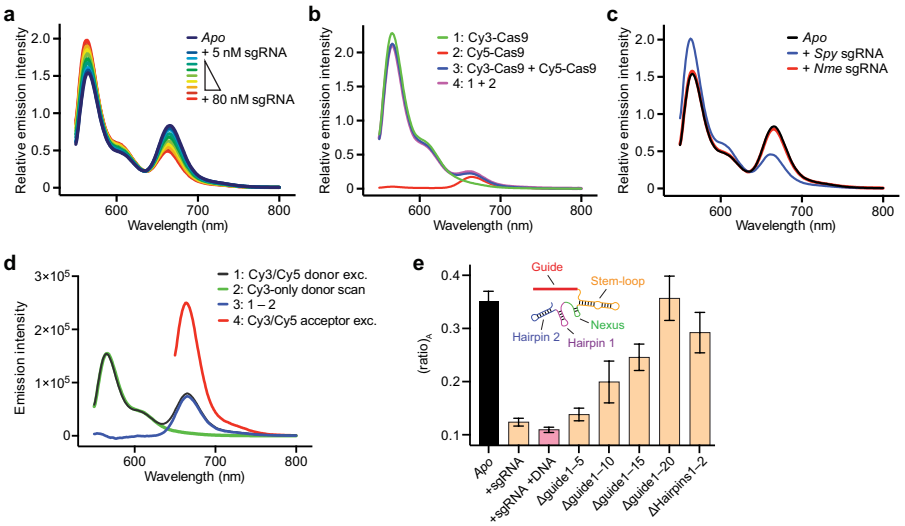
Experiments were replicated at least three times, and presented data are representative results unless stated otherwise.

Cas9 structural analysis. All structure figures were generated using Pymol (Schrödinger). Cas9 molecules from distinct crystal structures were aligned using the RuvC and PI domains (root mean squared deviation ≈0.5–0.7). To generate the modelled docked state for the HNH domain (Fig. 2a and Extended Data Fig. 3), nucleotides 12–13 of chain D of PDB 2QNC (endonuclease-VII-DNA structure) were first aligned to nucleotides 11–12 of chain C of PDB 4UN3 (sgRNA/DNA-bound Cas9 structure). A copy of the Cas9 HNH domain from PDB 4UN3 was then aligned to chain A of PDB 2QNC. Conservation rendering was done using a multiple sequence alignment of 250 Cas9 homologues and the ConSurf server[31].

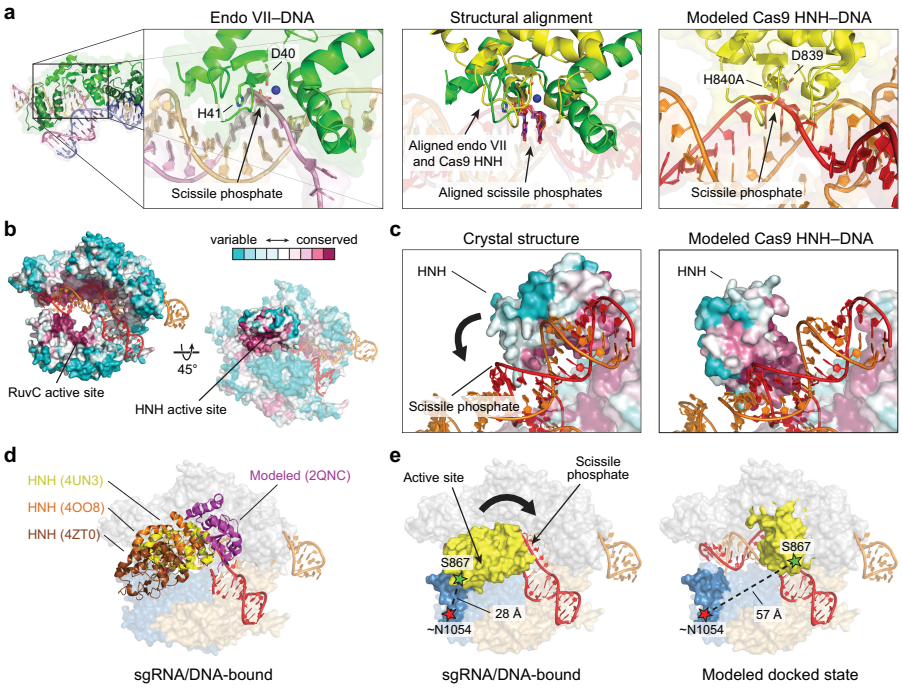
30. Sternberg, S. H., Haurwitz, R. E. & Doudna, J. A. Mechanism of substrate selection by a highly specific CRISPR endoribonuclease. *RNA* **18,** 661–672 (2012).
31. Ashkenazy, H., Erez, E., Martz, E., Pupko, T. & Ben-Tal, N. ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. *Nucleic Acids Res.* **38,** W529–W533 (2010).



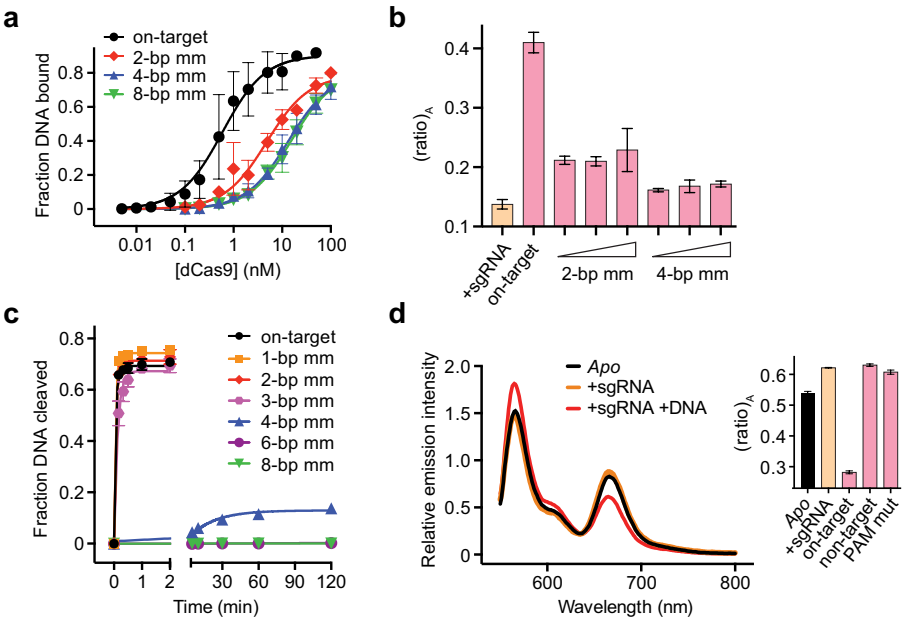
Extended Data Figure 1 | Biochemical preparation and DNA cleavage activity of dye-labelled Cas9. a, Size-exclusion chromatograms of Cy3/Cy5-labelling reactions with cysteine-free Cas9 (C80S/C574S) or the two double-cysteine Cas9 variants used to generate $Cas9_{hinge}$ and $Cas9_{HNH-1}$. Reactions contained 10 μM Cas9 and 200 μM Cy3- and Cy5-maleimide, and were separated on a Superdex 200 10/300 column (GE Healthcare). Cysteine-free Cas9 was unreactive. **b**, Sodium dodecyl sulphate–polyacrylamide gel electrophoresis (SDS–PAGE) analysis of unlabelled and dye-labelled Cas9 variants. The gel was scanned for Cy3 and Cy5 fluorescence (right) before being stained with Coomassie blue (left). For gel source data, see Supplementary Fig. 1. **c**, Representative radiolabelled DNA cleavage assay with wild-type (WT) Cas9 and doubly labelled Cas9 variants used in this study, resolved by denaturing PAGE (left); quantified data and exponential fits are shown on the right. S, substrate; NT, cleaved non-target strand; T, cleaved target strand. Error bars, s.d.; $n = 3$.



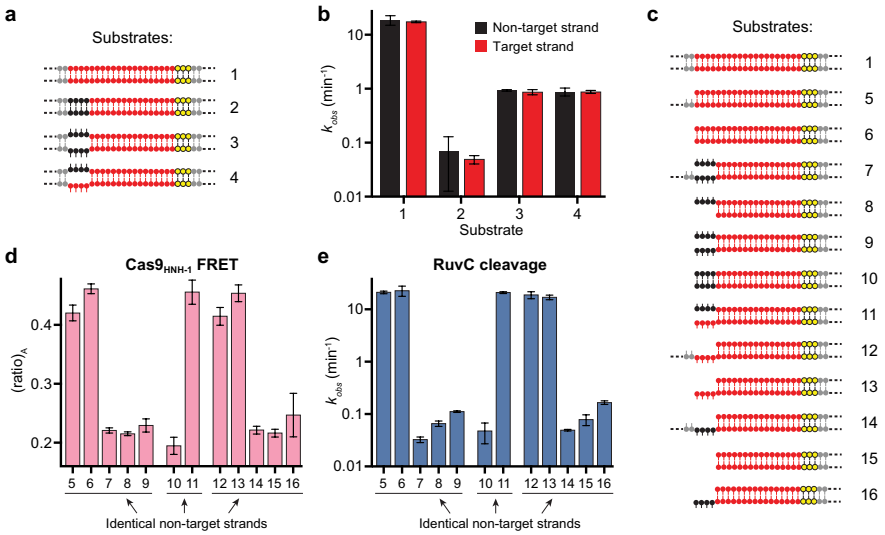
Extended Data Figure 2 | Fluorescence control experiments with Cas9$_{hinge}$ and dCas9$_{hinge}$, and representative analysis of fluorescence emission spectra to calculate (ratio)$_A$. **a**, Fluorescence emission spectra of 50 nM Cas9$_{hinge}$ in the presence of increasing concentrations of full-length sgRNA. Protein and sgRNA concentrations were calculated under non-denaturing conditions using theoretical extinction coefficients. **b**, Fluorescence emission spectra of (1) Cy3-labelled Cas9$_{hinge}$, (2) Cy5-labelled Cas9$_{hinge}$, and (3) an equal mixture of Cy3-Cas9$_{hinge}$ and Cy5-Cas9$_{hinge}$ upon excitation at 530 nm. The minor fluorescence peak for Cy5 in the mixed sample results from residual absorbance of Cy5-Cas9$_{hinge}$ at 530 nm and not from intermolecular FRET (compare spectra 3 with 4, which is a sum of spectra 1 and 2). **c**, Fluorescence emission spectra of Cas9$_{hinge}$ in the presence of sgRNA substrates specific to *S. pyogenes* (*Spy*) or *N. meningitidis* (*Nme*) Cas9. **d**, Determination of the (ratio)$_A$ parameter, which is proportional to FRET efficiency. Shown for *apo*-Cas9$_{hinge}$ are (1) an emission spectrum of Cy3/Cy5-Cas9$_{hinge}$ upon excitation of the donor at 530 nm; (2) an emission spectrum of donor only Cy3-Cas9$_{hinge}$ upon excitation of the donor at 530 nm, normalized to 1; (3) the extracted fluorescence of the acceptor via energy transfer, obtained by subtracting 2 from 1; and (4) an emission spectrum of Cy3/Cy5-Cas9$_{hinge}$ upon direct excitation of the acceptor at 630 nm. (Ratio)$_A$ is calculated by dividing the integrated intensity (650–800 nm) of 3 by the integrated intensity of 4. **e**, (Ratio)$_A$ data for dCas9$_{hinge}$ in the presence of the same sgRNA substrates tested with nuclease-active Cas9$_{hinge}$ in Fig. 1e. Error bars, s.d.; $n = 3$.



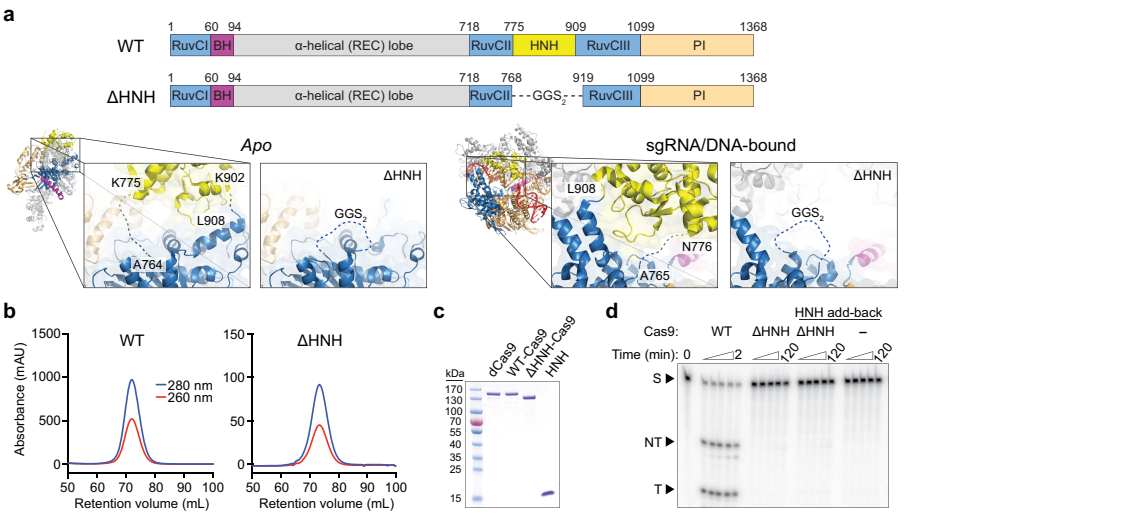
Extended Data Figure 3 | Modelling of the HNH domain docked at the cleavage site, and design of the Cas9$_{HNH-2}$ FRET construct. a, The scissile phosphate and flanking nucleotides of a DNA substrate co-crystallized with the phage T4 endonuclease VII (endo VII; PDB 2QNC; left) were aligned with the scissile phosphate and flanking nucleotides of the DNA target strand in the sgRNA/DNA-bound Cas9 crystal structure (PDB 4UN3; middle). Structural alignment of the Cas9 HNH domain with endonuclease VII (middle) results in a model of how the Cas9 HNH domain docks at the cleavage site (right). Catalytic residues are labelled, target strands are shown in red and pink, and a magnesium ion is depicted as a blue sphere. **b**, Conservation rendering of the sgRNA/DNA-bound Cas9 crystal structure, generated using ConSurf, shows that the most highly conserved patches of the HNH domain, including the active site, are solvent-exposed in the observed conformation. The HNH domain is omitted from the view on the left for clarity. **c**, Magnified view of the HNH domain in its observed conformation (left) and the model for the docked state (right), coloured as in **b**. The DNA target strand fits snugly in a groove on the HNH domain in the model, with the most highly conserved patches located in the immediate vicinity of the scissile phosphate. DNA and sgRNA are coloured red and orange, respectively. **d**, The conformational flexibility of the HNH domain in available Cas9 crystal structures is revealed by structural alignment of the nuclease lobe (RuvC and PI domains) from two sgRNA/DNA-bound structures (PDB accession numbers 4UN3 and 4OO8) and the sgRNA-bound structure (PDB 4ZT0). The modelled docked state from **a** is shown. **e**, Design of Cas9$_{HNH-2}$ FRET construct. Measured distances between ~N1054 and S867 in the sgRNA/DNA-bound Cas9 structure and a model of the HNH domain docked at the cleavage site are indicated. Putative conformational changes of the HNH domain are shown with a black arrow.



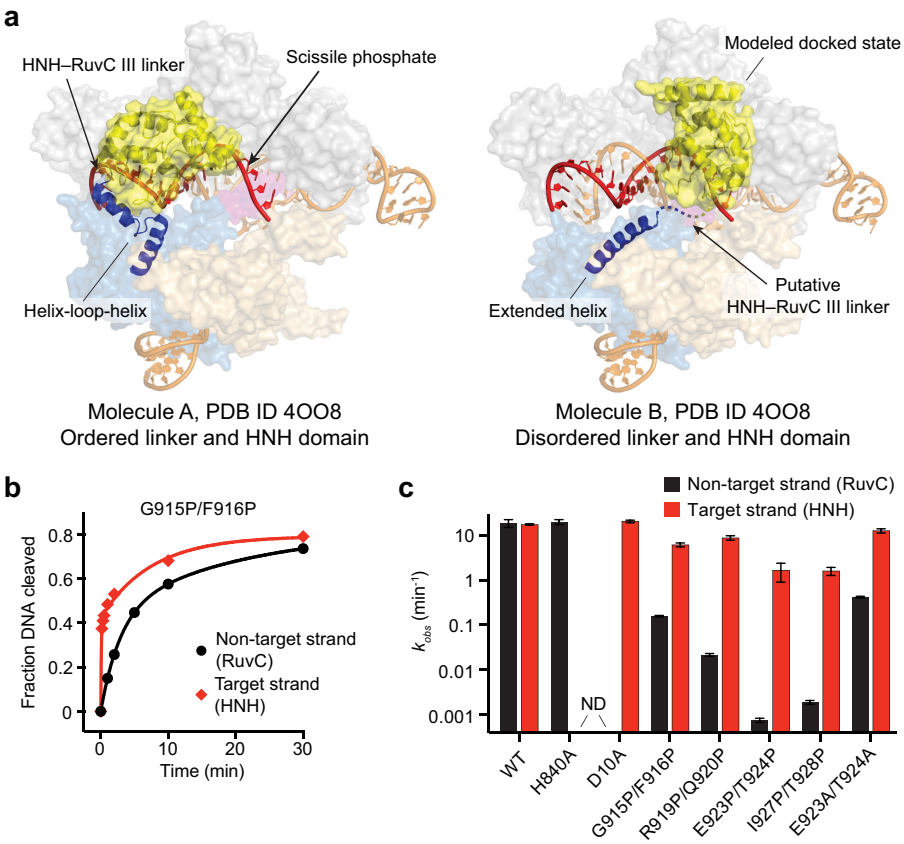
Extended Data Figure 4 | Evidence that variable (ratio)$_A$ values for dCas9$_{HNH-1}$ reflect distinct conformational states/dynamics, and FRET data for Cas9$_{HNH-2}$. a, DNA binding assay with dCas9 and either on-target DNA or off-target DNAs containing 2, 4, or 8-bp mismatches at the PAM-distal end. Binding fits are shown as solid lines and yield equilibrium dissociation constants (K_d) of 0.80, 6.7, 19, and 20 nM, respectively. Given these values, 99%, 96%, 89%, and 89% of dCas9 should be bound to DNA under the conditions used for FRET experiments in Fig. 2c (50 nM dCas9$_{HNH-1}$, 200 nM DNA). **b**, (Ratio)$_A$ data for 50 nM dCas9$_{HNH-1}$ in the presence of 1 μM sgRNA and either 200 nM, 400 nM, or 1 μM off-target DNAs containing 2- or 4-bp mismatches. Data for sgRNA only and on-target DNA are shown for comparison. **c**, DNA cleavage time courses for the indicated DNA substrates using wild-type Cas9. Exponential fits are shown as solid lines, and extracted rate constants are shown in Fig. 2d. **d**, Fluorescence emission spectra of Cas9$_{HNH-2}$ in the presence of the indicated substrates. The inset shows (ratio)$_A$ values; mut, mutation. Error bars in **a** and **b**–**d**, s.d.; $n = 3$–5 and 3, respectively.



Extended Data Figure 5 | Additional experimental support for dependence of RuvC nuclease activity on HNH conformational changes. **a**, Panel of DNA substrates tested in **b**, with on-target (1) at top. Matched and mismatched positions of DNA target strand sequences relative to the sgRNA are coloured red and black, respectively, with the PAM in yellow. Some substrates contain internal mismatches between the two DNA strands; dashed lines indicate additional flanking sequence. **b**, Kinetics of non-target (black) and target (red) strand cleavage for the indicated DNA substrates. **c**, Panel of DNA substrates tested in **d** and **e**, depicted as in **a**. **d**, $(Ratio)_A$ data for $Cas9_{HNH-1}$ in the presence of the indicated DNA substrates. **e**, Non-target strand cleavage kinetics of the RuvC domain for the indicated DNA substrates. Error bars in **b**, **d**, **e**, s.d.; $n = 3$.



Extended Data Figure 6 | Design, purification, and DNA cleavage activity of ΔHNH-Cas9. a, Domain organization of WT- and ΔHNH-Cas9 (top), showing the residues that were replaced with a GGS_2 linker to generate ΔHNH-Cas9. Magnified view of connections between the HNH domain and RuvC II and III motifs in the *apo* (left) and sgRNA/DNA-bound (right) Cas9 crystal structures, as well as in the ΔHNH-Cas9 construct. Disordered linkers and the introduced GGS_2 linker are shown as dashed lines. **b**, Size-exclusion chromatograms of WT- and ΔHNH-Cas9 using a Superdex 200 16/60 column (GE Healthcare). **c**, SDS–PAGE analysis of dCas9 (D10A/H840A), WT-Cas9, ΔHNH-Cas9, and the purified HNH domain (residues 776–907). Expected molecular masses are 159 kDa, 159 kDa, 142 kDa, and 16 kDa, respectively. For gel source data, see Supplementary Fig. 1. **d**, Representative radiolabelled DNA cleavage assay with WT-Cas9, ΔHNH-Cas9, ΔHNH-Cas9 in the presence of excess HNH domain, and HNH domain alone, resolved by denaturing PAGE.



Extended Data Figure 7 | Structural analysis and perturbation of the HNH–RuvC III linker. a, Molecules A (left) and B (right) of the sgRNA/DNA-bound Cas9 crystal structure (PDB 4OO8). Molecule A has an ordered HNH domain and HNH–RuvC III linker, whereas these are both disordered in molecule B; the missing density for the HNH domain is replaced with the modelled docked state (right). Another prominent difference is the N-terminal region of the RuvC III motif (blue helices), which rearranges from a helix–loop–helix in molecule A into an extended helix in molecule B. Proline pairs were inserted to prevent formation of this extended helix. **b**, Target (red) and non-target (black) strand cleavage time courses with the indicated Cas9 variant. Exponential fits are shown as solid lines. **c**, Kinetics of target (red) and non-target (black) strand cleavage for the indicated Cas9 mutants. ND, cleavage not detected. Error bars in **b** and **c**, s.d.; $n = 3$.

Extended Data Table 1 | Measured distances between residues labelled with FRET pairs

	Inter-residue distance *		
Structure used	D435–E945	S355–S867	S867–N1054
Apo (4CMP)	21 Å	79 Å	6 Å
sgRNA-bound (4ZT0)	78 Å	81 Å	7 Å
sgRNA/DNA-bound (4OO8 mol A)	77 Å †	61 Å	34 Å ‡
sgRNA/DNA-bound (4UN3)	83 Å	59 Å	28 Å §
sgRNA/DNA-bound, HNH docked state ‖		21 Å	57 Å §

*Distances were measured between Cα atoms of the indicated residues, except where indicated, for the denoted structures (PDB accession numbers in parentheses).
†E945 is disordered in the structure; an average of measured distances to T941 and I950 is reported.
‡N1054 is disordered in the structure; an average of measured distances to T1048 and I1063 is reported.
§N1054 is disordered in the structure; an average of measured distances to I1050 and K1059 is reported.
‖The docked state for the HNH domain was generated using PDB accession numbers 4UN3 and 2QNC.

Extended Data Table 2 | RNA and DNA substrates used in this study

Description	Sequence *
λ1-targeting sgRNA †	5′-GACGCAUAAAGAUGAGACGCGUUUUAGAGCUAUGCUGUUUUGGAAACAAAACAGCAUAGCAAGUUAAAAUAAGGCUAGUCCGUUAUCAACUUGAAAAAGUGGCACCGAGUCGGUGCUUUUUUUGGAUC-3′
λ1-targeting sgRNA, *Nme* Cas9 ‡	5′-GACUGACGCAUAAAGAUGAGACGCGUUGUAGCUCCCUUUCUCAUUUCGGAAACGAAAUGAGAACCGUUGCUACAAUAAGGCCGUCUGAAAAGAUGUGCCGCAACGCUCUGCCCCUUAAAGCUUCUGCUUUA AGGGGCAUCGUUUAUUGCUCGUGCGCUGGAUC-3′
λ1-targeting sgRNA, Δguide1–5	5′-GUAAAGAUGAGACGCGUUUUAGAGCUAUGCUGUUUUGGAAACAAAACAGCAUAGCAAGUUAAAAUAAGGCUAGUCCGUUAUCAACUUGAAAAAGUGGCACCGAGUCGGUGCUUUUUUUGGAUC-3′
λ1-targeting sgRNA, Δguide1–10	5′-GAUGAGACGCGUUUUAGAGCUAUGCUGUUUUGGAAACAAAACAGCAUAGCAAGUUAAAAUAAGGCUAGUCCGUUAUCAACUUGAAAAAGUGGCACCGAGUCGGUGCUUUUUUUGGAUC-3′
λ1-targeting sgRNA, Δguide1–15	5′-GACGCGUUUUAGAGCUAUGCUGUUUUGGAAACAAAACAGCAUAGCAAGUUAAAAUAAGGCUAGUCCGUUAUCAACUUGAAAAAGUGGCACCGAGUCGGUGCUUUUUUUGGAUC-3′
λ1-targeting sgRNA, Δguide1–20 §	5′-GGUUUUAGAGCUAUGCUGUUUUGGAAACAAAACAGCAUAGCAAGUUAAAAUAAGGCUAGUCCGUUAUCAACUUGAAAAAGUGGCACCGAGUCGGUGCUUUUUUUGGAUC-3′
λ1-targeting sgRNA, Δhairpin1	5′-GACGCAUAAAGAUGAGACGCGUUUUAGAGCUAUGCUGUUUUGGAAACAAAACAGCAUAGCAAGUUAAAAUAAGGCUAGUCCGUUAUCAACUUGAAAAAGUGGAUC-3′
λ1-targeting sgRNA, Δhairpins1–2	5′-GACGCAUAAAGAUGAGACGCGUUUUAGAGCUAUGCUGUUUUGGAAACAAAACAGCAUAGCAAGUUAAAAUAAGGCUAGUCCGUGGAUC-3′
λ1 on-target DNA, Substrate 1, Fig. 3	5′-AGCAGAAATCTCTGCTGACGCATAAAGATGAGACGCTGGAGTACAAACGTCAGCT-3′ 3′-TCGTCTTTAGAGACGACTGCGTATTTCTACTCTGCGACCTCATGTTTGCAGTCGA-5′
λ2 non-target DNA	5′-GAGTGGAAGGATGCCAGTGATAAGTGGAATGCCATGTGGGCTGTCAAAATTGAGC-3′ 3′-CTCACCTTCCTACGGTCACTATTCACCTTACGGTACACCCGACAGTTTTAACTCG-5′
λ1 off-target DNA, PAM mutation	5′-AGCAGAAATCTCTGCTGACGCATAAAGATGAGACGCTCGAGTACAAACGTCAGCT-3′ 3′-TCGTCTTTAGAGACGACTGCGTATTTCTACTCTGCGAGCTCATGTTTGCAGTCGA-5′
λ1 off-target DNA, seed mutation	5′-AGCAGAAATCTCTGCTGACGCATAAAGATGAGTGCGTGGAGTACAAACGTCAGCT-3′ 3′-TCGTCTTTAGAGACGACTGCGTATTTCTACTCACGCACCTCATGTTTGCAGTCGA-5′
λ1 off-target DNA, 1-bp mismatch	5′-AGCAGAAATCTCTGCTCACGCATAAAGATGAGACGCTGGAGTACAAACGTCAGCT-3′ 3′-TCGTCTTTAGAGACGAGTGCGTATTTCTACTCTGCGACCTCATGTTTGCAGTCGA-5′
λ1 off-target DNA, 2-bp mismatch	5′-AGCAGAAATCTCTGCTCTCGCATAAAGATGAGACGCTGGAGTACAAACGTCAGCT-3′ 3′-TCGTCTTTAGAGACGAGAGCGTATTTCTACTCTGCGACCTCATGTTTGCAGTCGA-5′
λ1 off-target DNA, 3-bp mismatch	5′-AGCAGAAATCTCTGCTCTGGCATAAAGATGAGACGCTGGAGTACAAACGTCAGCT-3′ 3′-TCGTCTTTAGAGACGAGACCGTATTTCTACTCTGCGACCTCATGTTTGCAGTCGA-5′
λ1 off-target DNA, 4-bp mismatch	5′-AGCAGAAATCTCTGCTCTGCCATAAAGATGAGACGCTGGAGTACAAACGTCAGCT-3′ 3′-TCGTCTTTAGAGACGAGACGGTATTTCTACTCTGCGACCTCATGTTTGCAGTCGA-5′
λ1 off-target DNA, 6-bp mismatch	5′-AGCAGAAATCTCTGCTCTGCGTTAAAGATGAGACGCTGGAGTACAAACGTCAGCT-3′ 3′-TCGTCTTTAGAGACGAGACGCAATTTCTACTCTGCGACCTCATGTTTGCAGTCGA-5′
λ1 off-target DNA, 8-bp mismatch	5′-AGCAGAAATCTCTGCTCTGCGTATAAGATGAGACGCTGGAGTACAAACGTCAGCT-3′ 3′-TCGTCTTTAGAGACGAGACGCATATTCTACTCTGCGACCTCATGTTTGCAGTCGA-5′
λ1 off-target DNA, Substrate 2, Fig. 3	5′-AGCAGAAATCTCTGCTCTGCCATAAAGATGAGACGCTGGAGTACAAACGTCAGCT-3′ 3′-TCGTCTTTAGAGACGAGACGGTATTTCTACTCTGCGACCTCATGTTTGCAGTCGA-5′
λ1 off-target DNA, Substrate 3, Fig. 3	5′-AGCAGAAATCTCTGCT^{GACG}CATAAAGATGAGACGCTGGAGTACAAACGTCAGCT-3′ 3′-TCGTCTTTAGAGACGAGACGGTATTTCTACTCTGCGACCTCATGTTTGCAGTCGA-5′
λ1 off-target DNA, Substrate 4, Fig. 3	5′-^{CTGC}CATAAAGATGAGACGCTGGAGTACAAACGTCAGCT-3′ 3′-CTGCGTATTTCTACTCTGCGACCTCATGTTTGCAGTCGA-5′
λ1 off-target DNA, Substrate 5, Fig. 3	5′-CATAAAGATGAGACGCTGGAGTACAAACGTCAGCT-3′ 3′-CTGCGTATTTCTACTCTGCGACCTCATGTTTGCAGTCGA-5′
λ1 off-target DNA, Substrate 6, Fig. 3	5′-^{GACG}CATAAAGATGAGACGCTGGAGTACAAACGTCAGCT-3′ 3′-GACGGTATTTCTACTCTGCGACCTCATGTTTGCAGTCGA-5′
λ1 off-target DNA, Substrate 7, Fig. 3	5′-CATAAAGATGAGACGCTGGAGTACAAACGTCAGCT-3′ 3′-GACGGTATTTCTACTCTGCGACCTCATGTTTGCAGTCGA-5′
λ1 off-target DNA, Substrate 8, Fig. 3	5′-AGCAGAAATCTCTGCTGACGCATAAAGATGAGAC^{GC}TGGAGTACAAACGTCAGCT-3′ 3′-TCGTCTTTAGAGACGACTGCGTATTTCTACTCTGGCACCTCATGTTTGCAGTCGA-5′
λ1 off-target DNA, Substrate 9, Fig. 3	5′-AGCAGAAATCTCTGCTGACGCATAAAGATGAG^{ACGC}TGGAGTACAAACGTCAGCT-3′ 3′-TCGTCTTTAGAGACGACTGCGTATTTCTACTCACGCACCTCATGTTTGCAGTCGA-5′

Description	Sequence *
λ1 off-target DNA, Substrate 2, ED Fig. 5	5′-AGCAGAAATCTCTGCTCTGCCATAAAGATGAGACGCTGGAGTACAAACGTCAGCT-3′ 3′-TCGTCTTTAGAGACGAGACGGTATTTCTACTCTGCGACCTCATGTTTGCAGTCGA-5′
λ1 off-target DNA, Substrate 3, ED Fig. 5	5′-AGCAGAAATCTCTGCT^{GACG}CATAAAGATGAGACGCTGGAGTACAAACGTCAGCT-3′ 3′-TCGTCTTTAGAGACGAGACGGTATTTCTACTCTGCGACCTCATGTTTGCAGTCGA-5′
λ1 off-target DNA, Substrate 4, ED Fig. 5	5′-AGCAGAAATCTCTGCT^{CTGC}CATAAAGATGAGACGCTGGAGTACAAACGTCAGCT-3′ 3′-TCGTCTTTAGAGACGACTGCGTATTTCTACTCTGCGACCTCATGTTTGCAGTCGA-5′
λ1 off-target DNA, Substrate 5, ED Fig. 5	5′-GACGCATAAAGATGAGACGCTGGAGTACAAACGTCAGCT-3′ 3′-TCGTCTTTAGAGACGACTGCGTATTTCTACTCTGCGACCTCATGTTTGCAGTCGA-5′
λ1 off-target DNA, Substrate 6, ED Fig. 5	5′-GACGCATAAAGATGAGACGCTGGAGTACAAACGTCAGCT-3′ 3′-CTGCGTATTTCTACTCTGCGACCTCATGTTTGCAGTCGA-5′
λ1 off-target DNA, Substrate 7, ED Fig. 5	5′-^{GACG}CATAAAGATGAGACGCTGGAGTACAAACGTCAGCT-3′ 3′-TCGTCTTTAGAGACGAGACGGTATTTCTACTCTGCGACCTCATGTTTGCAGTCGA-5′
λ1 off-target DNA, Substrate 8, ED Fig. 5	5′-^{GACG}CATAAAGATGAGACGCTGGAGTACAAACGTCAGCT-3′ 3′-GTATTTCTACTCTGCGACCTCATGTTTGCAGTCGA-5′
λ1 off-target DNA, Substrate 9, ED Fig. 5	5′-^{GACG}CATAAAGATGAGACGCTGGAGTACAAACGTCAGCT-3′ 3′-GACGGTATTTCTACTCTGCGACCTCATGTTTGCAGTCGA-5′
λ1 off-target DNA, Substrate 10, ED Fig. 5	5′-CTGCCATAAAGATGAGACGCTGGAGTACAAACGTCAGCT-3′ 3′-GACGGTATTTCTACTCTGCGACCTCATGTTTGCAGTCGA-5′
λ1 off-target DNA, Substrate 11, ED Fig. 5	5′-^{CTGC}CATAAAGATGAGACGCTGGAGTACAAACGTCAGCT-3′ 3′-CTGCGTATTTCTACTCTGCGACCTCATGTTTGCAGTCGA-5′
λ1 off-target DNA, Substrate 12, ED Fig. 5	5′-CATAAAGATGAGACGCTGGAGTACAAACGTCAGCT-3′ 3′-TCGTCTTTAGAGACGACTGCGTATTTCTACTCTGCGACCTCATGTTTGCAGTCGA-5′
λ1 off-target DNA, Substrate 13, ED Fig. 5	5′-CATAAAGATGAGACGCTGGAGTACAAACGTCAGCT-3′ 3′-CTGCGTATTTCTACTCTGCGACCTCATGTTTGCAGTCGA-5′
λ1 off-target DNA, Substrate 14, ED Fig. 5	5′-CATAAAGATGAGACGCTGGAGTACAAACGTCAGCT-3′ 3′-TCGTCTTTAGAGACGAGACGGTATTTCTACTCTGCGACCTCATGTTTGCAGTCGA-5′
λ1 off-target DNA, Substrate 15, ED Fig. 5	5′-CATAAAGATGAGACGCTGGAGTACAAACGTCAGCT-3′ 3′-GTATTTCTACTCTGCGACCTCATGTTTGCAGTCGA-5′
λ1 off-target DNA, Substrate 16, ED Fig. 5	5′-CATAAAGATGAGACGCTGGAGTACAAACGTCAGCT-3′ 3′-GACGGTATTTCTACTCTGCGACCTCATGTTTGCAGTCGA-5′

*sgRNA guide sequences and matching DNA target strand sequences are shown in red. PAM sites (5′-NGG-3′) are highlighted in yellow on the non-target strand. Internal mismatches in select DNA substrates are denoted by misaligned text on the non-target strand.
†All sgRNA constructs contain remnants of the BamHI sequence on the 3′ end resulting from run-off *in vitro* transcription.
‡sgRNA specific to *N. meningitidis* (*Nme*) Cas9 contains an additional 3′ extension, which does not affect activity (data not shown), for purposes unrelated to this study.
§Δguide1–20 sgRNA contains an extraneous 5′-G from *in vitro* transcription.

LETTER

doi:10.1038/nature15525

Crystal structure of the RNA-dependent RNA polymerase from influenza C virus

Narin Hengrung[1,2], Kamel El Omari[2], Itziar Serna Martin[1], Frank T. Vreede[1], Stephen Cusack[3], Robert P. Rambo[4], Clemens Vonrhein[5], Gérard Bricogne[5], David I. Stuart[2,4], Jonathan M. Grimes[2,4]§ & Ervin Fodor[1]§

Negative-sense RNA viruses, such as influenza, encode large, multidomain RNA-dependent RNA polymerases that can both transcribe and replicate the viral RNA genome[1]. In influenza virus, the polymerase (FluPol) is composed of three polypeptides: PB1, PB2 and PA/P3. PB1 houses the polymerase active site, whereas PB2 and PA/P3 contain, respectively, cap-binding and endonuclease domains required for transcription initiation by cap-snatching[2]. Replication occurs through *de novo* initiation and involves a complementary RNA intermediate. Currently available structures of the influenza A and B virus polymerases include promoter RNA (the 5′ and 3′ termini of viral genome segments), showing FluPol in transcription pre-initiation states[3,4]. Here we report the structure of apo-FluPol from an influenza C virus, solved by X-ray crystallography to 3.9 Å, revealing a new 'closed' conformation. The apo-FluPol forms a compact particle with PB1 at its centre, capped on one face by PB2 and clamped between the two globular domains of P3. Notably, this structure is radically different from those of promoter-bound FluPols[3,4]. The endonuclease domain of P3 and the domains within the carboxy-terminal two-thirds of PB2 are completely rearranged. The cap-binding site is occluded by PB2, resulting in a conformation that is incompatible with transcription initiation. Thus, our structure captures FluPol in a closed, transcription pre-activation state. This reveals the conformation of newly made apo-FluPol in an infected cell, but may also apply to FluPol in the context of a non-transcribing ribonucleoprotein complex. Comparison of the apo-FluPol structure with those of promoter-bound FluPols allows us to propose a mechanism for FluPol activation. Our study demonstrates the remarkable flexibility of influenza virus RNA polymerase, and aids our understanding of the mechanisms controlling transcription and genome replication.

FluPol is a highly flexible protein complex; however, the conformational states it can adopt are uncharacterized. Understanding the nature of these conformational states is central to determining the regulatory mechanisms of this enzyme. To this end, we have determined the structure of FluPol from influenza C virus[5] ($FluPol_C$), in the absence of promoter RNA. We expressed all three individual subunits of $FluPol_C$ in insect cells by infection with a single baculovirus construct. $FluPol_C$ purified from this system was active in both replication and transcription initiation (Extended Data Fig. 1). We crystallized apo-$FluPol_C$ in two different crystal forms (Extended Data Table 1), and solved its structure at 3.9 Å (Extended Data Fig. 2) and 4.3 Å resolution.

Our model of $FluPol_C$ (Fig. 1) comprises 711 of the 754 residues of PB1 (94.3%), 762 out of 774 for PB2 (98.4%) and 693 out of 709 for P3 (97.7%). $FluPol_C$ forms a relatively compact structure (Fig. 1a, b). P3 folds into two domains connected by a long linker (Fig. 1c): an amino-terminal endonuclease domain ($P3_{endo}$) and a C-terminal domain ($P3_C$), which sandwiches PB1 at the heart of the molecule. PB1 has the canonical right-hand-like polymerase fold, possessing palm, fingers and thumb subdomains with additional N- and C-terminal extensions ($PB1_{N\text{-}ext}$ and $PB1_{C\text{-}ext}$) that facilitate interactions with the other subunits (Fig. 1d). The thumb of PB1 is reinforced by $P3_C$. The priming loop of PB1, believed to facilitate *de novo* replication initiation[4], is not visible in our structure and is probably disordered. PB2 stacks against one face of PB1, contacting both domains of P3. PB2 comprises 9 domains: the N-terminal PB1 interaction domain ($PB2_{N\text{-}ter}$), $PB2_{N1}$, $PB2_{N2}$, $PB2_{lid}$ and $PB2_{mid}$ domains, a cap-binding domain ($PB2_{cap}$), a linker domain ($PB2_{cap\text{-}627\ linker}$), the 627 domain ($PB2_{627}$) and a C-terminal nuclear localization signal (NLS) ($PB2_{NLS}$) domain (Fig. 1e).

The fold of each $FluPol_C$ domain is very similar to its counterpart in $FluPol_A$ and $FluPol_B$, even though the sequence identity between these polymerases is only ~30% (Extended Data Table 2 and Supplementary Fig. 1). The average root mean squared deviation (r.m.s.d.) values of Cα atoms between equivalent superposed domains of $FluPol_C$ and $FluPol_A$, or of $FluPol_C$ and $FluPol_B$ are 1.6 Å or 1.5 Å, respectively, demonstrating that the FluPol fold is conserved across influenza A, B and C viruses. All key active site residues within $FluPol_C$ are structurally conserved, and we confirmed, by mutation, that $FluPol_C$ shares common mechanisms with $FluPol_A$ (Extended Data Fig. 3a). The PB1 subunits of FluPol A, B and C belong to a structural grouping that most closely resembles the polymerases of Reoviridae and Cystoviridae/Flaviridae (Extended Data Fig. 3b).

However, there are substantial differences between apo-$FluPol_C$ and the activated structures for promoter-bound $FluPol_A$ and $FluPol_B$. Most striking are the position of $P3_{endo}$ and the arrangement of the C-terminal domains of PB2 (Fig. 2, Supplementary Video 1 and Extended Data Table 3). Thus, $PB2_{627}$, which in $FluPol_A$ houses a crucial polymorphism (Glu627Lys) for the determination of viral host range and pathogenicity[6], lies level with the endonuclease domain in the apo structure, (Fig. 2a), whereas in the activated structures it lies close to $PB1_{palm}$ (Fig. 2b). The $PB2_{mid}$ and $PB2_{cap\text{-}627\ linker}$ domains are rearranged *en bloc*, by a rotation of 140° and a translation of 30 Å, between the apo and activated conformations. $PB2_{cap}$ also changes; in the apo structure, it is tucked in between the $PB1_{palm}$ and $PB2_{cap\text{-}627\ linker}$, while in the activated structures it does not extensively contact other domains. Finally, $PB2_{NLS}$ packs between $PB1_{C\text{-}ter}$ helix α23 and $P3_{endo}$ helix α7 in apo-$FluPol_C$ (Fig. 3a), but is near the base of the $PB1_{palm}$ in the activated structures (Fig. 2b). The movement of $PB2_{NLS}$ amounts to a rotation of 130° and a translation of 90 Å. The regions rearranged within PB2 match those that are disordered in the FluB2 structure[4], lying immediately downstream of a conserved glycine (PB2 residue 255 in $FluPol_C$).

The buried area between PB2 and P3 (5,000 $Å^2$) is more extensive in the apo than the activated, promoter-bound structures, reflecting new contacts between PB2 and $P3_{endo}$ (Fig. 3a and Extended Data Fig. 4a, b). Additionally, the extreme C terminus of PB2 is visible in our apo

[1]Sir William Dunn School of Pathology, University of Oxford, South Parks Road, Oxford OX1 3RE, UK. [2]Division of Structural Biology, Henry Wellcome Building for Genomic Medicine, University of Oxford, Oxford OX3 7BN, UK. [3]European Molecular Biology Laboratory, Grenoble Outstation and University Grenoble Alpes-Centre National de la Recherche Scientifique-EMBL Unit of Virus Host-Cell Interactions, 71 Avenue des Martyrs, CS 90181, 38042 Grenoble Cedex 9, France. [4]Diamond Light Source Ltd, Harwell Science & Innovation Campus, Didcot OX11 0DE, UK. [5]Global Phasing Ltd, Sheraton House, Castle Park, Cambridge CB3 0AX, UK.
§These authors jointly supervised this work.

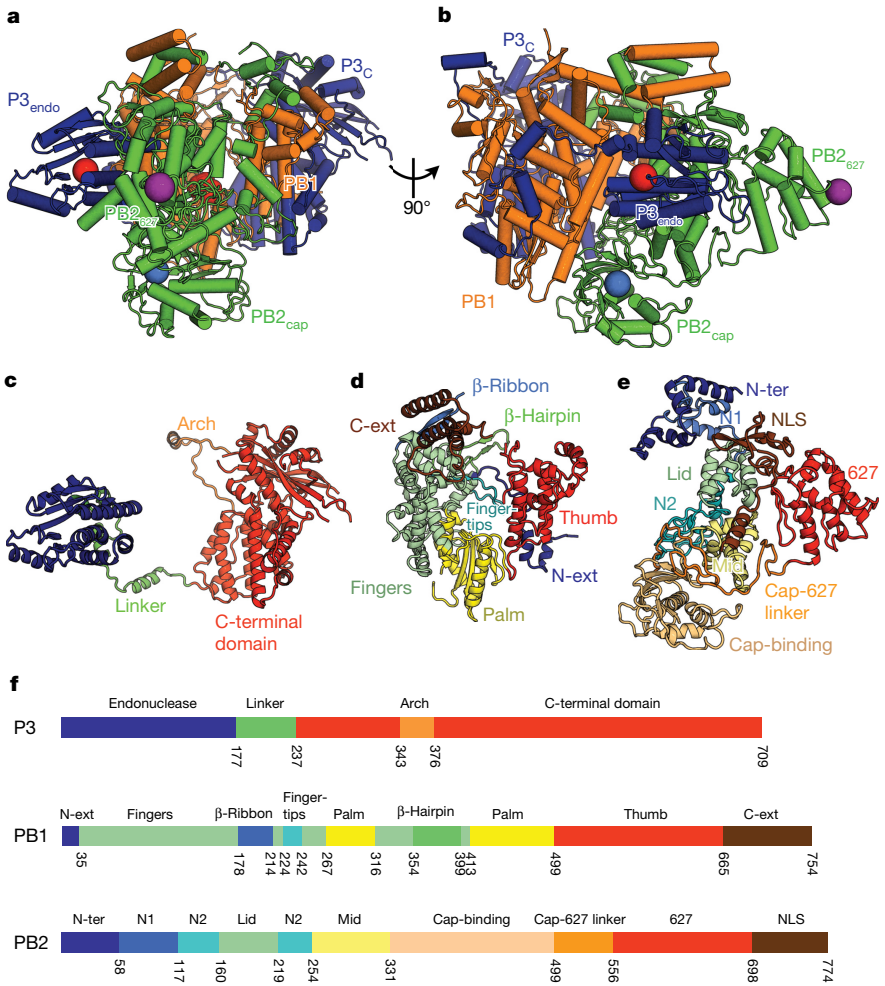


Figure 1 | Structure of FluPol$_C$. **a**, **b**, Two views of the structure of the FluPol$_C$ heterotrimer, coloured according to subunit (PB1, orange; PB2, green; P3, blue). The cap-binding pocket and endonuclease active site are shown as blue and red spheres, respectively. In **a**, the PB1 catalytic aspartates, residues 446 and 447, are also highlighted red. The position of PB2 residue 649 (equivalent to PB2 residue 627 in FluPol$_A$) is marked by a purple sphere. **c**–**e**, Structures of FluPol$_C$ subunits P3 (**c**), PB1 (**d**) and PB2 (**e**), coloured and labelled by domain. **f**, Domain maps of each FluPol$_C$ subunit.

structure (Extended Data Fig. 4b), forming a helix (α30) that packs against the back face of P3$_{endo}$ (near P3 helices α2, α3 and α7). There are also many new contacts between PB2 and PB1 (Fig. 2 and Extended Data Fig. 4c).

One important consequence of the arrangement seen in apo-FluPol$_C$ is that the cap-binding pocket is occluded (Fig. 3b). PB2$_{cap}$ is folded in on the rest of the subunit, facing residues 520–535 from the PB2$_{cap\text{-}627\ linker}$ domain. This is consistent with the observation that promoter RNA is required for FluPol cap-binding and endonuclease activity[7,8]. Thus, the structure we observe represents a closed, pre-activation state of FluPol and suggests that the viral RNA (vRNA) promoter causes the rearrangements necessary to form the activated structure.

Alternatively, the structure of FluPol$_C$ could indicate a fundamental conformational difference between FluPols from different influenza viruses. To clarify this, we performed small angle X-ray scattering (SAXS) experiments with FluPol$_C$, as these allowed us to distinguish between closed and activated FluPol conformations (Extended Data Fig. 5a). The observed scattering profiles from FluPol$_C$ were similar to a profile calculated from the FluPol$_A$ crystal structure, indicating that FluPol$_C$ can adopt the same activated conformation (Extended Data Fig. 5b). Thus, the change we see in the FluPol$_C$ crystal is not an influenza virus type difference. However, promoter RNA was not required for the activated conformation to be detected, indicating that changes between apo and promoter-bound structures need not exclusively be caused by RNA binding. This suggests that the energy barrier between different FluPol conformations is low. Indeed, when placed into a phosphate-based buffer, FluPol$_C$ adopted a currently uncharacterized conformation that was even more open than that of the promoter-bound structures (Extended Data Fig. 5c). These results suggest that FluPol may be poised between several different conformations, with only subtle environmental changes needed for a particular conformation to be favoured.

In line with this assessment, differences around the promoter-binding site between the apo-FluPol$_C$ and promoter-bound structures are small. Minor changes are evident around the pocket that binds the intra-base paired hook structure of the 5′ strand of the vRNA promoter (Extended Data Fig. 6); however, sequence alignments suggest that these differences are influenza-virus-type-specific. More interesting are the differences around the binding site for the 3′ strand of the vRNA promoter. In the apo structure, PB2 helix α4, PB2$_{N1}$ and the associated region of PB1$_{thumb}$ lie ~5 Å further away from the polymerase core than in the promoter-bound structures (Fig. 3c). This change is transmitted to the neighbouring PB1$_{C\text{-}ext}$–PB2$_{N\text{-}ter}$ interaction domain through PB1 helix α22, resulting in a 20° rotation of this domain between the apo and vRNA promoter-bound conformations (Fig. 3d). Since the PB1$_{C\text{-}ext}$–PB2$_{N\text{-}ter}$ domain lies next to PB2$_{NLS}$ in apo-FluPol$_C$ (Fig. 3a), this rotation could trigger the movement of PB2$_{NLS}$ from its apo position, leading to the subsequent massive reorganization of FluPol after vRNA–promoter binding (Supplementary Video 2).

Notably, only one currently reported FluPol structure (FluB2) contains a fully ordered vRNA promoter (in the others, the 3′ vRNA strand is either truncated or partially disordered)[4]. However, this does not

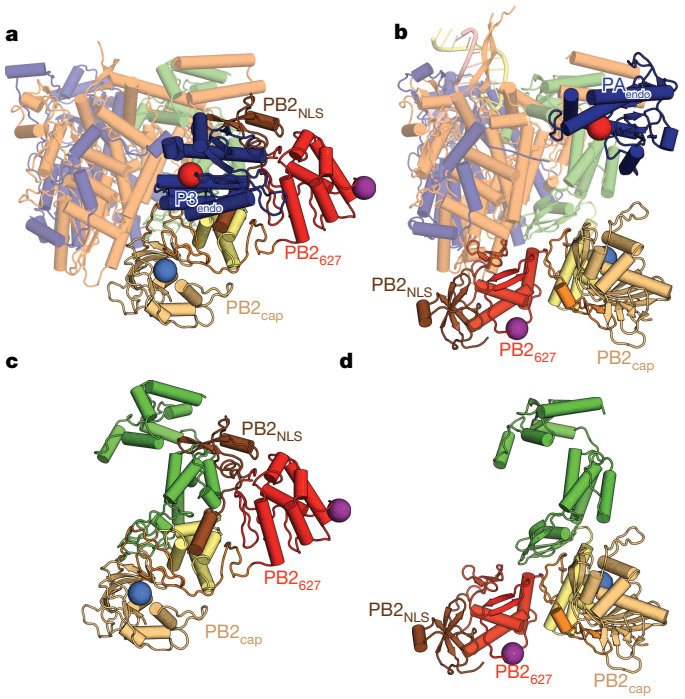


Figure 2 | Comparison of apo-FluPol$_C$ with promoter-bound FluPol$_A$. **a**, Apo-FluPol$_C$, depicted in the same orientation and colouring as in Fig. 1b, but with the C-terminal domains of PB2 coloured as in Fig. 1e. Domains that do not change between apo and promoter-bound conformations are depicted as semi-transparent. **b**, Promoter-bound FluPol$_A$, shown as in **a**. **c**, **d**, The PB2 subunits of apo-FluPol$_C$ (**c**) and promoter-bound FluPol$_A$ (**d**), depicted as in **a** and **b**.

display a stable activated conformation, as the C-terminal two-thirds of PB2 are not resolved. We suggest that this is because initial binding of the vRNA promoter, into a resting position away from the active site, generates a dynamic equilibrium between closed and activated conformations. The activated structure is only seen when the 3′ end of the 3′ vRNA promoter strand is either not present or disordered[3,4]. Hence, the activated conformation might only be fully stabilized when this 3′ end is released from its resting position to enter the polymerase active site. To test this hypothesis, we compared the ability of a full-length or truncated (lacking four nucleotides at the 3′ end of the 3′ strand) vRNA promoter to stimulate FluPol$_C$ cap-dependent cleavage activity (Fig. 4). We reasoned that stabilization of an activated over a closed conformation would enhance capped-RNA cleavage, as the cap-binding pocket in PB2 becomes more accessible. In line with this, we observed a significant enhancement in capped RNA cleavage in the presence of the truncated promoter RNA (Fig. 4). The relative inefficiency of the full-length vRNA promoter to stimulate cleavage supports our assertion that initial promoter binding results in a closed/activated equilibrium. The mechanism behind this may involve the PB1 β-ribbon (177–212 in FluPol$_A$)[3], which is disordered in the apo-FluPol$_C$ structure, but adopts different conformations in the activated and FluB2 structures.

In summary, we have solved the structure of the RNA polymerase from an influenza C virus in the absence of RNA, uncovering a closed conformation accessible to FluPol. Our structure explains the observation that FluPol in the absence of promoter RNA is unable to perform cap-snatching[7,8], and we propose a mechanism for how vRNA promoter might bring about FluPol activation. However, the closed conformation captured here may have a wider functional relevance, because it could still be accessible to FluPol bound to a fully ordered vRNA promoter that does not enter the active site. Therefore, in the context of a non-transcribing viral ribonucleoprotein complex (RNP), containing FluPol, RNA and nucleoprotein, FluPol may well adopt this closed conformation. In addition, dependent on stabilization of the PB1 priming loop, the closed conformation that we observe might still allow *de novo* initiation, as this is not dependent on cap-snatching. Thus, the conformation that we observe, in addition to being a transcription pre-activation state, could be relevant during genome replication initiation. This would allow the activity of FluPol within an RNP to be regulated by other viral factors and host proteins[9–12].

Our work underlines the tremendous flexibility of this protein complex. This flexibility offers an explanation for the differences between several low-resolution electron microscopy reconstructions of RNP-associated FluPol, as well as explaining why the promoter-bound structures do not fit well into these reconstructions[13–16]. Furthermore, since

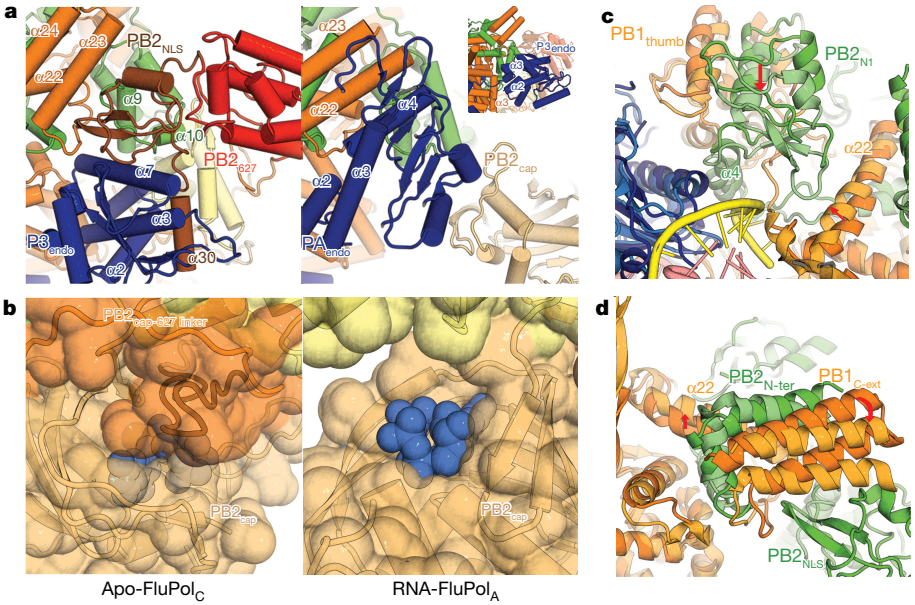


Figure 3 | Critical changes between apo-FluPol$_C$ and promoter-bound FluPol$_A$. **a**, Equivalent views of FluPol$_C$ (left) and FluPol$_A$ (right), showing domain arrangement differences. The inset shows the arrangement of P3$_{endo}$ within the FluPol$_C$ structure. **b**, Close-up of the FluPol$_C$ (left) and FluPol$_A$ (right) cap-binding domains. PB2 residues 520–535 in FluPol$_C$ are coloured dark orange. The cap-binding pocket is shown with blue spheres. **c**, **d**, Two views of a superposition of apo-FluPol$_C$ and FluPol$_A$, with FluPol$_C$ coloured as in Fig. 1 and FluPol$_A$ in lighter colours. 5′ and 3′ promoter RNAs in the FluPol$_A$ structure are coloured pink and yellow, respectively. Arrows highlight differences between the two conformations.

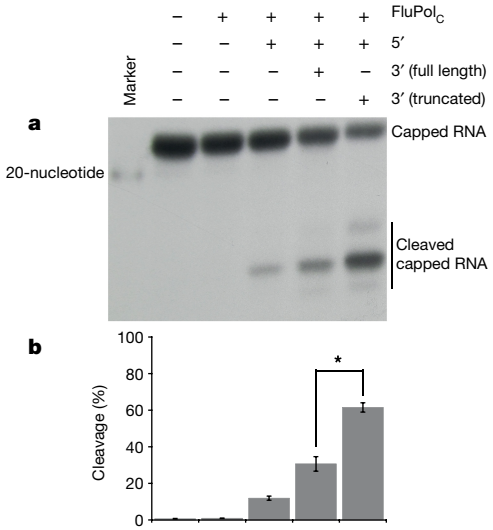


Figure 4 | Capped-RNA cleavage assays with FluPol$_C$. a, Representative autoradiograph of a capped-RNA cleavage assay. In each reaction, radiolabelled capped RNA was incubated at 30 °C for 2 h with FluPol$_C$ and the indicated strands of the vRNA promoter. **b**, Quantification of cleavage, expressed as the percentage of cleaved to total RNA, from three replicates of this assay, performed with the same polymerase preparation. Mean cleavage percentage is plotted. Error bars show s.d. Asterisk indicates a significant difference between cleavage with full or truncated vRNA promoter ($n = 3$, $P = 0.0003$, two-tailed t-test).

negative-sense RNA virus polymerases share a common organization, with a central polymerase core surrounded by various functional modular appendages[1,17], the conformational flexibility revealed here might be a theme among all of these polymerases and not just particular to FluPol[18].

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 20 September 2014; accepted 25 August 2015.
Published online 26 October 2015.

1. Ortín, J. & Martin-Benito, J. The RNA synthesis machinery of negative-stranded RNA viruses. *Virology* **479–480,** 532–544 (2015).
2. Fodor, E. The RNA polymerase of influenza a virus: mechanisms of viral transcription and replication. *Acta Virol.* **57,** 113–122 (2013).
3. Pflug, A., Guilligay, D., Reich, S. & Cusack, S. Structure of influenza A polymerase bound to the viral RNA promoter. *Nature* **516,** 355–360 (2014).
4. Reich, S. *et al.* Structural insight into cap-snatching and RNA synthesis by influenza polymerase. *Nature* **516,** 361–366 (2014).
5. Muraki, Y. & Hongo, S. The molecular virology and reverse genetics of influenza C virus. *Jpn. J. Infect. Dis.* **63,** 157–165 (2010).
6. Gabriel, G. & Fodor, E. Molecular determinants of pathogenicity in the polymerase complex. *Curr. Top. Microbiol. Immunol.* **385,** 35–60 (2014).
7. Cianci, C., Tiley, L. & Krystal, M. Differential activation of the influenza virus polymerase via template RNA binding. *J. Virol.* **69,** 3995–3999 (1995).
8. Rao, P., Yuan, W. & Krug, R. M. Crucial role of CA cleavage sites in the cap-snatching mechanism for initiating viral mRNA synthesis. *EMBO J.* **22,** 1188–1198 (2003).
9. Chang, S. *et al.* Cryo-EM structure of influenza virus RNA polymerase complex at 4.3 Å resolution. *Mol. Cell* **57,** 925–935 (2015).
10. Kawaguchi, A. & Nagata, K. De novo replication of the influenza virus RNA genome is regulated by DNA replicative helicase, MCM. *EMBO J.* **26,** 4566–4575 (2007).
11. Paterson, D. & Fodor, E. Emerging roles for the influenza A virus nuclear export protein (NEP). *PLoS Pathog.* **8,** e1003019 (2012).
12. York, A., Hengrung, N., Vreede, F. T., Huiskonen, J. T. & Fodor, E. Isolation and characterization of the positive-sense replicative intermediate of a negative-strand RNA virus. *Proc. Natl Acad. Sci. USA* **110,** E4238–E4245 (2013).
13. Area, E. *et al.* 3D structure of the influenza virus polymerase complex: Localization of subunit domains. *Proc. Natl Acad. Sci. USA* **101,** 308–313 (2004).
14. Arranz, R. *et al.* The structure of native influenza virion ribonucleoproteins. *Science* **338,** 1634–1637 (2012).
15. Coloma, R. *et al.* The structure of a biologically active influenza virus ribonucleoprotein complex. *PLoS Pathog.* **5,** e1000491 (2009).
16. Moeller, A., Kirchdoerfer, R. N., Potter, C. S., Carragher, B. & Wilson, I. A. Organization of the influenza virus replication machinery. *Science* **338,** 1631–1634 (2012).
17. Morin, B., Kranzusch, P. J., Rahmeh, A. A. & Whelan, S. P. J. The polymerase of negative-stranded RNA viruses. *Curr. Opin. Virol.* **3,** 103–110 (2013).
18. Gerlach, P., Malet, H., Cusack, S. & Reguera, J. Structural insights into Bunyavirus replication and its regulation by the vRNA promoter. *Cell* **161,** 1267–1279 (2015).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank N. Naffakh for providing plasmids and I. Berger for providing us with the MultiBac system. The authors thank Diamond Light Source for beamtime (proposal mx8423), and the staff of the MX beamlines for assistance. K. Harlos also provided valuable technical assistance. We thank G. G. Brownlee for initiating this project and for his comments and continuous encouragement, and members of the Fodor, Grimes and Stuart laboratories, particularly A. te Velthuis, for discussions. This work was supported by Medical Research Council (MRC) grants MR/K000241/1 (to E.F.), G1100138 (to F.T.V.) and G1000099 (to D.I.S.), Wellcome Trust Studentship 092931/Z/10/Z (to N.H.), an MRC Studentship (to I.S.M.), and Wellcome Trust administrative support grant 075491/Z/04.

Author Contributions N.H., F.T.V., D.I.S., J.M.G. and E.F. conceived and designed the study. N.H., K.E.O., I.S.M. and R.P.R. performed experiments. S.C. provided coordinates before publication and discussed the results. C.V. and G.B. aided with data analysis and provided valuable advice. N.H., F.T.V., K.E.O., R.P.R., D.I.S., J.M.G. and E.F. analysed data and wrote the paper.

Author Information Atomic coordinates and structure factors for the reported crystal structure have been deposited in the Protein Data Bank under accession numbers 5D98 and 5D9A. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to J.M.G. (jonathan@strubi.ox.ac.uk) or E.F. (ervin.fodor@path.ox.ac.uk).

METHODS

Protein expression and purification. The three subunits of the influenza C/Johannesburg/1/1966 virus polymerase (PB1: AAF89738, PB2: AAF89739, P3: AAF89737) were co-expressed in Sf9 cells from codon-optimized genes (GeneArt) cloned into a single baculovirus using the MultiBac system[19]. Expression and purification of $FluPol_C$ proceeded as previously described for $FluPol_A$[12], except that the gel filtration buffer used was 0.5 M NaCl, 25 mM HEPES-NaOH, pH 7.5 and 10% (v/v) glycerol. For crystallization and storage, protein purified in this buffer was supplemented with 0.5 mM TCEP, and 10 mM $MgCl_2$ or 10 mM $CaCl_2$.

Crystallization, data collection and structure determination. Crystals of $FluPol_C$, belonging to two different space groups, grew from sitting-drop vapour-diffusion experiments at 20 °C[20,21], set up using a protein:precipitant ratio of 2:1. In these experiments, 5 mg ml^{-1} protein was mixed with either 70% (v/v) Morpheus G2 (Molecular Dimensions), supplemented with 0%–1% 1 M NaOH, to generate $P4_32_12$ crystals; or with crystal-seeds and 0.2 M NaCl, 0.1 M Na-HEPES, pH 7.5 and 25% (w/v) PEG 4000, for $P2_12_12_1$ crystals. For heavy atom derivatization, $P4_32_12$ crystals were soaked in a solution of gold(I) potassium cyanide dissolved in mother liquor, for 2–3 h at 20 °C. Crystals were cryo-protected using 25% (v/v) glycerol in crystallization buffer, before flash-cooling in liquid nitrogen, and data collection on beamlines I03 and I04 at the Diamond Light Source, Didcot, UK. The beam size was matched to the crystal size and data were collected on a Pilatus 6M detector at a wavelength of 0.9763 Å (tetragonal native), 1.0350 Å (tetragonal derivative) and 0.9795 Å (orthorhombic native). Data collection statistics are shown in Extended Data Table 1. Data were processed using Xia2 (ref. 22) and HKL2000 (ref. 23). Initial phases were obtained by single isomorphous replacement with anomalous scattering (SIRAS), using data from native $P4_32_12$ and gold-derivatized crystals. The $P4_32_12$ data used at this stage was collected earlier (at a wavelength of 0.8634 Å) than that subsequently used in refinement. Heavy atoms were located with SHELX[24] and phases improved by two-fold non-crystallographic averaging (the crystallographic asymmetric unit contained two heterotrimers) and solvent flattening (solvent content 76%) using Phenix.autosol[25]. The tetragonal and the orthorhombic data were sharpened to 40 $Å^2$ and 36 $Å^2$, respectively. The $P2_12_12_1$ crystals were solved by molecular replacement (program Phaser[26]), using the $P4_32_12$ structure as the search model. As expected the orthorhombic crystals possessed four heterotrimers in the crystallographic asymmetric unit, allowing phase improvement using non-crystallographic symmetry (NCS) averaging and solvent flattening using general averaging program (GAP) (D.I.S. and J.M.G., unpublished observations). The published fragments of $FluPol_A$ (PDB accessions 4IUJ, 4AWH, 4CB4, 3A1G and 2VY7) were fitted by eye using Coot, which was used for all model building[27]. Comparison with the complete $FluPol_A$ and $FluPol_B$ structures (4WSB and 4WSA, respectively), aided by the anomalous scattering from the sulphur atoms as markers, allowed us to build and refine complete models for $FluPol_C$. This provided a total of six independent views of the polymerase. Performing superpositions of these demonstrated that the molecule adopts a virtually identical conformation across all copies from both crystal forms (mean pairwise r.m.s.d. in Cα was 0.94 Å between all pairs of molecules across both space groups). Refinement (Extended Data Table 1) used BUSTER[28] aided by NCS and initially phase restraints, and REFMAC[29] with secondary structure restraints using PROSMART[30].

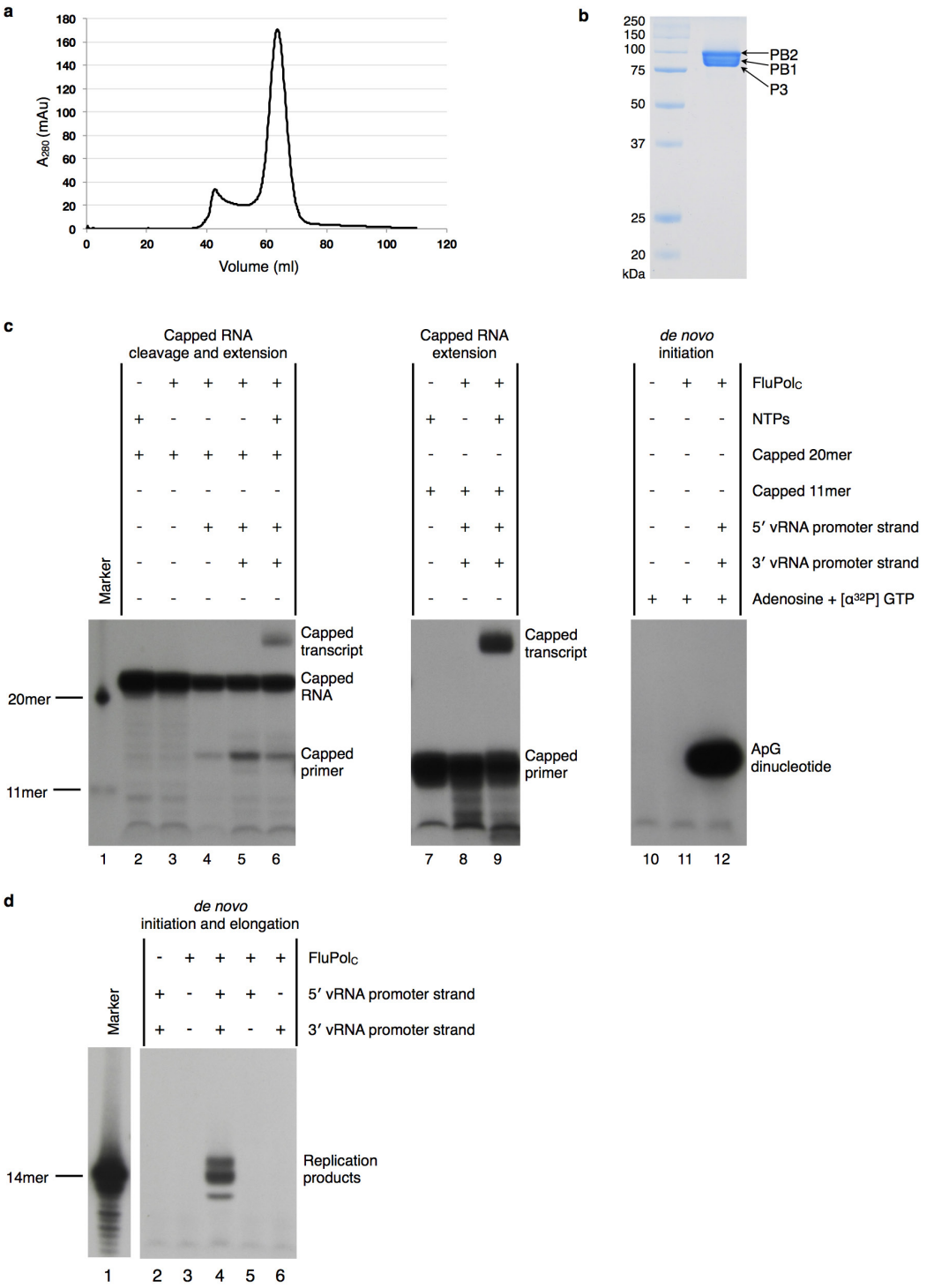
SAXS experiments. SAXS measurements were performed on beamline B21 at Diamond Light Source, Didcot, UK. Samples were prepared onsite using a Shodex Kw-403 size exclusion column and Agilent HPLC. Approximately 40–60 µl of sample were collected for SAXS at 20 °C using a sample to detector distance of 3.9 m and X-ray wavelength of 1 Å. Samples were exposed for 300 s in 10 s acquisition blocks. Images were corrected for variations in beam current, normalized for exposure time and processed into 1D scattering curves using GDA and DAWN. Buffer subtractions and all other subsequent analysis were performed with the program ScÅtter (http://www.bioisis.net/scatter). Samples were checked for radiation damage by visual inspection of the Guinier region as a function of exposure time.

Data analysis. Figures and videos were prepared using PyMOL (http://www.pymol.org) and Chimera[31]. Structural comparisons used SHP[32].

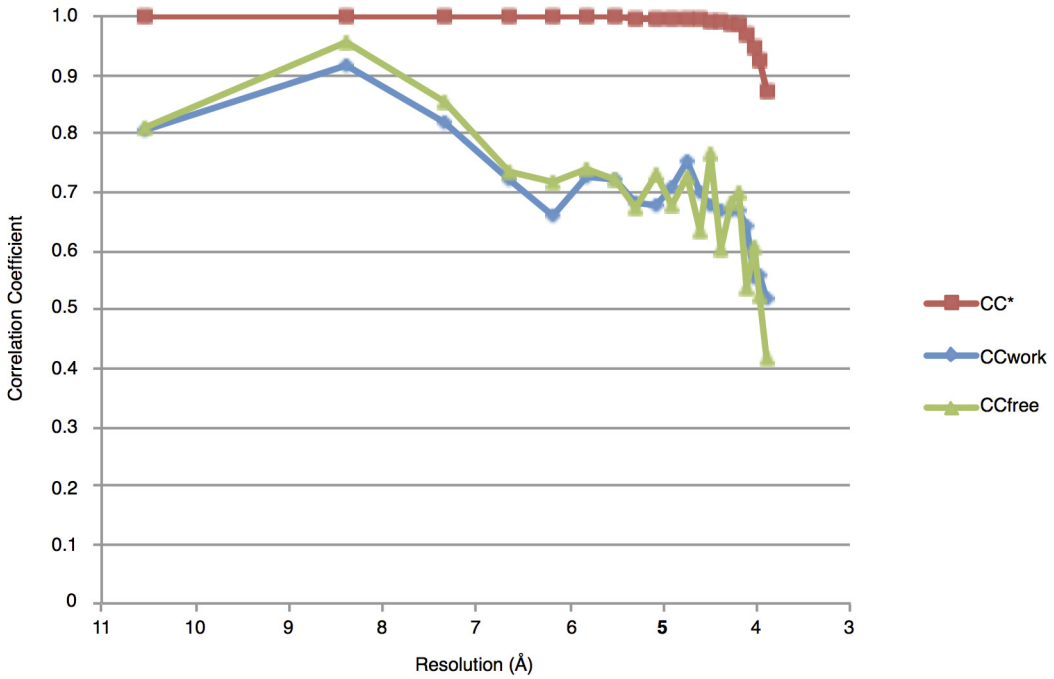
Polymerase activity assays. For the cap-dependent cleavage and transcription assays, $FluPol_C$ (400 ng per reaction) was incubated for 2 h at 30 °C with or without (as indicated) NTPs (1 mM ATP, 0.5 mM each CTP/UTP/GTP), radiolabelled capped 20-nucleotide or 11-nucleotide RNAs, 0.6 µM each 5′ and 3′ vRNA promoter strands, in a reaction buffer containing 7.5 mM $MgCl_2$, 1.0 mM TCEP, 2 U $µl^{-1}$ RNasin (Promega), 20 mM HEPES-NaOH, pH 7.5, 100 mM NaCl and 5% (v/v) glycerol. For the *de novo* initiation and elongation assays, $FluPol_C$ (400 or 800 ng per reaction, as indicated) was incubated for 2–3 h at 30 °C with 2.5 mM adenosine and 0.075 µM [α-^{32}P]GTP or 1 mM ATP, 0.5 mM each CTP/UTP, 0.1 mM GTP, 0.3 µM [α-^{32}P]GTP and (as indicated) 0.6 µM each 5′ or 3′ vRNA promoter strands, in the same reaction buffer as above. The reaction volume was 4 µl for all reactions. Products were denatured by boiling (98 °C, 5 min) after the addition of formamide (4 µl) and separated on a denaturing 20% polyacrylamide gel, with the indicated size markers. Products were visualized by autoradiography. For all activity assays except the cleavage assays, the sequences of the promoter RNA oligonucleotides used were: 5′-AGCAGUAGCAAGGAG-3′ (5′ vRNA) and 5′-CUCCUGCUUCUGCU-3′ (3′ vRNA). The sequences of the RNAs used in the capped-RNA cleavage assays were 5′-AGCAGUAGCAAGGGG-3′ (5′), 5′-UAUACCCCUGCUUC-3′ (3′ truncated) or 5′-UAUACCCCUGCUUCUGCU-3′ (3′ full length).

Capped and radiolabelled RNA was produced by incubating 5′ diphosphate synthetic 20-nucleotide (5′-ppAAUCUAUAAUAGCAUUAUCC-3′)[3,4] or 11-nucleotide (5′-ppGAAUACUCAAG-3′)[33,34] RNA (Chemgenes), with [α-^{32}P]GTP, vaccinia virus capping enzyme (NEB) and 2′-*O*-methyltransferase (NEB), following the manufacturer's instructions. The resulting RNAs were gel purified before use in the above assays.

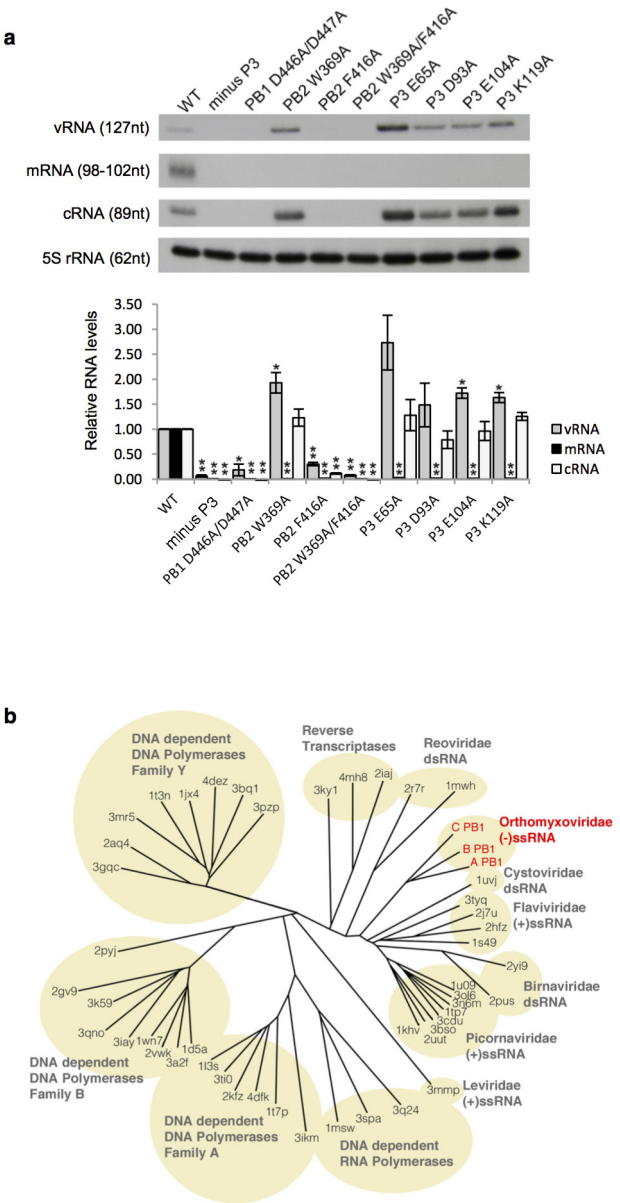
19. Bieniossek, C., Imasaki, T., Takagi, Y. & Berger, I. MultiBac: expanding the research toolbox for multiprotein complexes. *Trends Biochem. Sci.* **37,** 49–57 (2012).
20. Walter, T. S. *et al.* A procedure for setting up high-throughput nanolitre crystallization experiments. I. Protocol design and validation. *J. Appl. Crystallogr.* **36,** 308–314 (2003).
21. Walter, T. S. *et al.* A procedure for setting up high-throughput nanolitre crystallization experiments. Crystallization workflow for initial screening, automated storage, imaging and optimization. *Acta Crystallogr. D* **61,** 651–657 (2005).
22. Winter, G., Lobley, C. M. C. & Prince, S. M. Decision making in xia2. *Acta Crystallogr. D* **69,** 1260–1273 (2013).
23. Otwinowski, Z. & Minor, W. Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol.* **276,** 307–326 (1997).
24. Sheldrick, G. M. Experimental phasing with SHELXC/D/E: combining chain tracing with density modification. *Acta Crystallogr. D* **66,** 479–485 (2010).
25. Adams, P. D. *et al.* PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D* 66, 213–221 (2010).
26. Mccoy, A. J. *et al.* Phaser crystallographic software. *J. Appl. Crystallogr.* **40,** 658–674 (2007).
27. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr. D* **60,** 2126–2132 (2004).
28. Bricogne, G. *et al.* BUSTER version 2.11.5 (Global Phasing Ltd., 2011).
29. Murshudov, G. N. *et al.* REFMAC5 for the refinement of macromolecular crystal structures. *Acta Crystallogr. D* **67,** 355–367 (2011).
30. Nicholls, R. A., Long, F. & Murshudov, G. N. Low-resolution refinement tools in REFMAC5. *Acta Crystallogr. D* **68,** 404–417 (2012).
31. Pettersen, E. F. *et al.* UCSF Chimera–a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25,** 1605–1612 (2004).
32. Stuart, D. I., Levine, M., Muirhead, H. & Stammers, D. K. Crystal-structure of cat muscle pyruvate-kinase at a resolution of 2.6 A. J. *Mol. Biol.* **134,** 109–142 (1979).
33. Brownlee, G. G., Fodor, E., Pritlove, D. C., Gould, K. G. & Dalluge, J. J. Solid phase synthesis of 5′-diphosphorylated oligoribonucleotides and their conversion to capped m7Gppp-oligoribonucleotides for use as primers for influenza A virus RNA polymerase in vitro. *Nucleic Acids Res.* **23,** 2641–2647 (1995).
34. Chung, T. D. *et al.* Biochemical studies on capped RNA primers identify a class of oligonucleotide inhibitors of the influenza virus RNA polymerase. *Proc. Natl Acad. Sci. USA* **91,** 2372–2376 (1994).
35. Deng, T., Sharps, J. L. & Brownlee, G. G. Role of the influenza virus heterotrimeric RNA polymerase complex in the initiation of replication. *J. Gen. Virol.* **87,** 3373–3377 (2006).
36. Crescenzo-Chaigne, B., Naffakh, N. & van der Werf, S. Comparative analysis of the ability of the polymerase complexes of influenza viruses type A, B and C to assemble into functional RNPs that allow expression and replication of heterotypic model RNA templates *in vivo*. *Virology* **265,** 342–353 (1999).
37. Fodor, E. *et al.* A single amino acid mutation in the PA subunit of the influenza virus RNA polymerase inhibits endonucleolytic cleavage of capped RNAs. *J. Virol.* **76,** 8989–9001 (2002).
38. Biswas, S. K. & Nayak, D. P. Mutational analysis of the conserved motifs of influenza A virus polymerase basic protein 1. *J. Virol.* **68,** 1819–1826 (1994).
39. Dias, A. *et al.* The cap-snatching endonuclease of influenza virus polymerase resides in the PA subunit. *Nature* **458,** 914–918 (2009).
40. Guilligay, D. *et al.* The structural basis for cap binding by influenza virus polymerase subunit PB2. *Nature Struct. Mol. Biol.* **15,** 500–506 (2008).
41. Yuan, P. *et al.* Crystal structure of an avian influenza polymerase PA_N reveals an endonuclease active site. *Nature* **458,** 909–913 (2009).
42. Fechter, P. *et al.* Two aromatic residues in the PB2 subunit of influenza A RNA polymerase are crucial for cap binding. *J. Biol. Chem.* **278,** 20381–20388 (2003).
43. Krissinel, E. & Henrick, K. Inference of macromolecular assemblies from crystalline state. *J. Mol. Biol.* **372,** 774–797 (2007).
44. Karplus, P. A. & Diederichs, K. Linking crystallographic model and data quality. *Science* **336,** 1030–1033 (2012).
45. Rice, P., Longden, I. & Bleasby, A. EMBOSS: The European molecular biology open software suite. *Trends Genet.* **16,** 276–277 (2000).



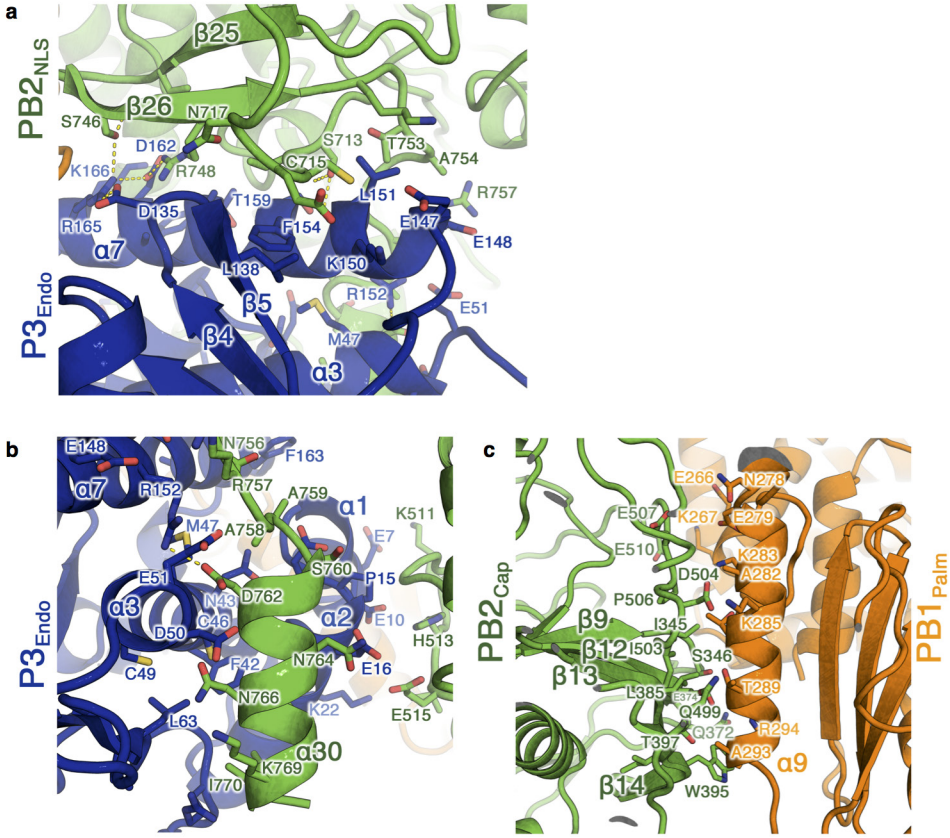
Extended Data Figure 1 | Purification and characterization of FluPol$_C$. **a**, Elution profile of FluPol$_C$, after affinity purification over IgG-sepharose, from a size-exclusion chromatography column. Eluted protein was detected by measuring the absorbance at 280 nm. **b**, Fractions corresponding to the major peak eluting from the size-exclusion chromatography column were mixed and analysed by SDS–PAGE on a 15% polyacrylamide gel, alongside the indicated molecular mass markers. Protein was visualized by Coomassie blue staining (PB1, 86.0 kDa; PB2, 87.2 kDa; P3, 81.9 kDa). **c**, Transcription and replication initiation assays. Lanes 2–6 test for transcription initiation. With the addition of vRNA promoter only (lanes 4 and 5), FluPol$_C$ can cleave a capped and radiolabelled 20-nucleotide RNA, demonstrating promoter-dependent endonuclease activity. Lane 6 shows that with the addition of NTPs, this capped primer can be extended to produce a capped transcript, thus demonstrating transcription initiation activity. This result is confirmed by lanes 7–9, which test for extension of a capped and radiolabelled 11-nucleotide RNA primer. Extension only takes place when the polymerase is supplied with NTPs and promoter RNA (lane 9). Lanes 10–12 assay for replication initiation. Lane 12 shows that FluPol$_C$ (400 ng per reaction) is able to synthesize ApG dinucleotide in a primer-independent manner. This demonstrates *de novo* replication initiation activity. Uncapped 20-nucleotide and 11-nucleotide primers are used as size markers in lane 1. The slow migration of the ApG dinucleotide compared to the markers is due to the lack of phosphate groups on the 5′ end of this product[35]. **d**, *De novo* initiation and elongation assay. FluPol$_C$ (800 ng) was incubated for 3 h with NTPs, [α-^{32}P]GTP and 5′ or 3′ vRNA promoter strands, as indicated. In the presence of both promoter strands (lane 4), FluPol$_C$ is able to produce a full-length copy of the template (14 nucleotides, corresponding to the major band), demonstrating *de novo* replication initiation and elongation activity. The minor slower and faster bands may correspond to non-templated extension and premature termination products, respectively.



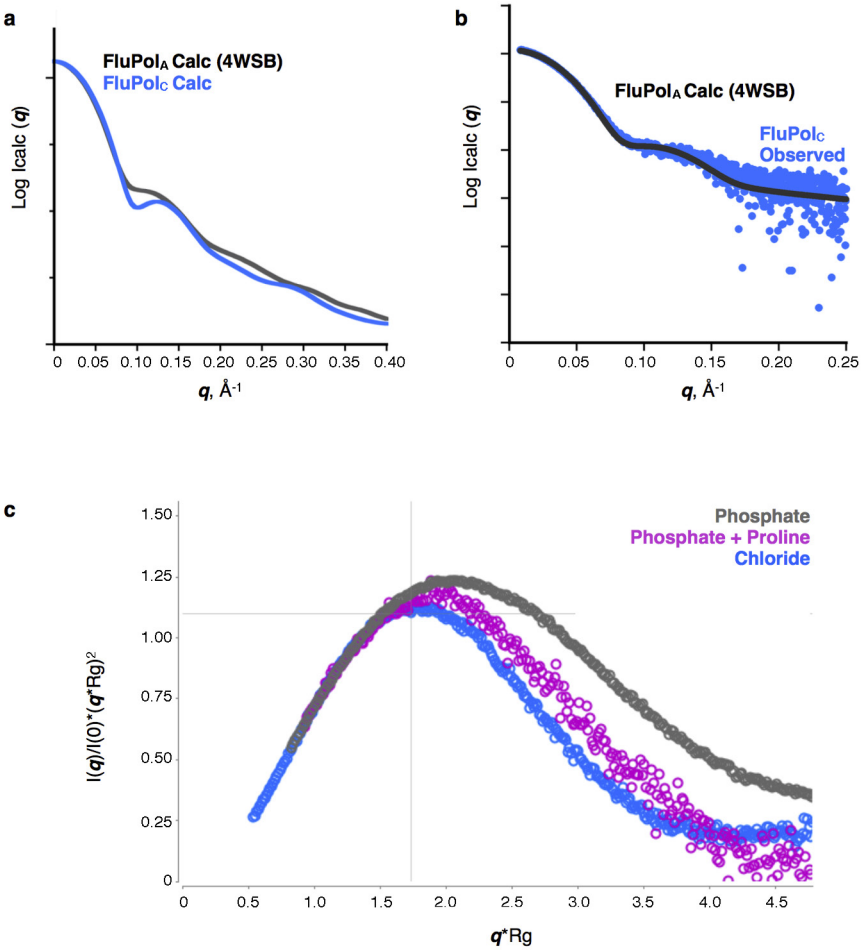
Extended Data Figure 2 | Data and model quality. Plots of CC*, CC_{free} and CC_{work} against resolution, for the tetragonal crystal data set and model. The CC* statistic assesses the signal present in the data, while CC_{free} and CC_{work} provide an estimate of the agreement between data and model. A CC* value of 0.87 for the highest resolution shell indicates that these data contain useful information up to 3.9 Å. CC_{free} and CC_{work} are lower than CC*, showing that the model does not overfit the data.



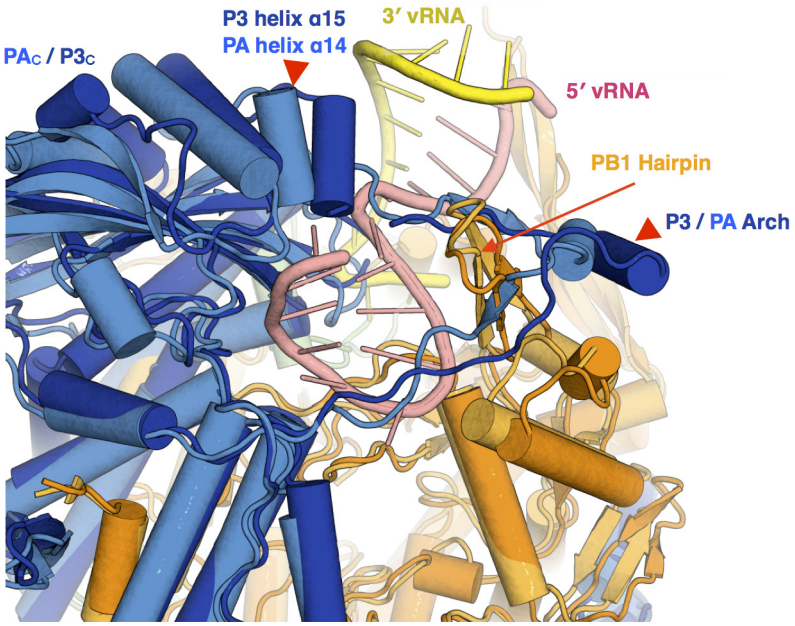
Extended Data Figure 3 | Functional and structural relationships of FluPol$_C$. a, Effect of amino acid mutations in FluPol$_C$ on transcription and replication. Plasmids to express FluPol$_C$ subunits and NP along with a plasmid expressing a negative-sense CAT reporter gene flanked by the terminal non-coding sequences of the influenza C virus NS gene segment[36] were transfected into 293T cells (ATCC). Total RNA was isolated using Trizol (Invitrogen) 30 h post transfection and viral RNAs were analysed by primer extension[37] using the following primers: 5′-CGCAAGGCGACAAGGTGCTGA-3′ (for detection of vRNA, yielding a 127-nucleotide product) and 5′-ATGTTCTTTACGATGCGATTGGG-3′ (for detection of mRNA and complementary RNA (cRNA), yielding 98–102-nucleotide and 89-nucleotide products, respectively). Primer extension products were analysed by 6% PAGE. Quantification of primer extension analysis using phosphorimaging is shown below. The mean and s.d. of three experiments are shown. Asterisks indicate a significant difference from wild type (WT), which was set to 100% (*$P < 0.05$; **$P < 0.01$, based on a two-sample t-test). A double mutation of two aspartic acids Asp446/Asp447, that align with Asp445/Asp446 of influenza A virus PB1 found to be critical for activity[38], resulted in no detectable activity in the context of FluPol$_C$. Mutation of amino acid residues in the PB2 cap-binding and P3 endonuclease domains that align with critical amino acid residues in FluPol$_A$[39–41] resulted in undetectable accumulation of viral mRNA although most of these mutants were still able to replicate. The exception was Phe416Ala in PB2 that inhibited both transcription and replication, suggesting that this mutation might not only affect cap-binding but overall PB2 folding, in agreement with previously observed inhibitory effects of mutations in the cap-binding domain on replication[42]. The requirement of these amino acid residues for mRNA synthesis is consistent with the hypothesis that FluPol$_C$ generates capped RNA primers for transcription initiation in a manner similar to that of FluPol$_A$. **b**, Structure-based phylogenetic tree showing the relationship of PB1 from FluPol$_C$ (C PB1) to other right-handed polymerases. Pairwise comparisons were performed using SHP[32] and a phylogenetic tree constructed using PHYLIP. The branches are identified by the PDB accession code of the polymerase.



Extended Data Figure 4 | New subunit interfaces in apo-FluPol$_C$. a, **b**, Two views of the interaction interface between PB2$_{NLS}$ (green) and P3$_{endo}$ (blue). Predicted polar contacts between the subunits are shown as dotted yellow lines. **c**, The interface between PB2$_{cap}$ (green) and PB1$_{palm}$ (orange). In all panels, residues at the interface were calculated using the 'Protein interfaces, surfaces and assemblies' service PISA at the European Bioinformatics Institute (http://www.ebi.ac.uk/pdbe/prot_int/pistart.html)[43].



Extended Data Figure 5 | SAXS analysis of FluPol$_C$. a, Calculated solution-state SAXS profiles for the crystal structures of vRNA-FluPol$_A$[3] (activated conformation) and apo-FluPol$_C$ (closed conformation). A distinguishing difference between these profiles is the dip at $q \sim 0.1$ Å^{-1} in the FluPol$_C$ curve. **b**, Solution-state SAXS profile of FluPol$_C$, without added promoter RNA, overlaid with the calculated curve for the vRNA-FluPol$_A$ structure[3]. The good match between these curves suggests that in this particular buffer (0.5 M NaCl, 25 mM HEPES-NaOH, pH 7.5, 5% (v/v) glycerol), FluPol$_C$ adopts the same globular conformation as the RNA bound state. **c**, Dimensionless Kratky plot of apo-FluPol$_C$ in the presence of 0.5 M NaCl, 25 mM HEPES-NaOH, pH 7.5 and 5% (v/v) glycerol (blue) or 100 mM KCl, 2% (w/v) sucrose and 100 mM sodium phosphate, pH 7.3, with (magenta) or without (grey) 200 mM proline. Cross-hairs denote the Guinier–Kratky point (1.732, 1.104), the peak position for an ideal, globular particle. As indicated by the upward shift of the peaks in the dimensionless Kratky plot, FluPol$_C$ is less globular in the presence of phosphate than it is in the 0.5 M NaCl buffer. This effect can be lessened if proline is also present, potentially owing to increased molecular crowding.



Extended Data Figure 6 | Differences at the 5′ vRNA promoter binding site. Superposition of apo-FluPol$_C$ (darker colours) and FluPol$_A$ (lighter colours) structures, with sites of interest labelled.

Extended Data Table 1 | Data collection, phasing and refinement statistics

	Tetragonal Native*	Tetragonal Derivative*	Orthorhombic Native*
Data collection			
Space group	$P4_32_12$	$P4_32_12$	$P2_12_12_1$
Cell dimensions			
a, *b*, *c* (Å)	185.66, 185.66, 598.22	184.22, 184.22, 598.75	107.28, 217.50, 597.75
α, β, γ (°)	90.00, 90.00, 90.00	90.00, 90.00, 90.00	90.00, 90.00, 90.00
Resolution (Å) **	100.6 – 3.9 (4.0 – 3.9)	127.3– 6.9 (7.1 – 6.9)	80.9– 4.3 (4.4 – 4.3)
R_{sym} or R_{merge}	0.196 (3.586)	0.157 (0.665)	0.204 (0.993)
$I/\sigma I$	13.2 (1.3)	16.5 (1.8)	5.5 (1.1)
Completeness (%)	98.8 (96.9)	99.1 (89.0)	99.0 (93.1)
Redundancy	21.2 (20.6)	14.5 (3.7)	3.3 (2.9)
$CC_{1/2}$ ***	(0.621)	(0.612)	(0.500)
Refinement			
Resolution (Å)	50.0 – 3.9		50.0 – 4.3
No. reflections	90,335		90,390
R_{work}/ R_{free}	0.286/0.326		0.316/0.368
No. atoms			
Protein	34,720		69,371
Wilson B-factors (Å^2)			
Protein	205		190
R.m.s deviations			
Bond lengths (Å)	0.008		0.009
Bond angles (°)	1.12		1.24

*The native data sets were each collected from a single crystal, whereas the derivative data set was produced by merging data from two crystals.
*Highest resolution shell is shown in parentheses.
**As described in ref. 44.

Extended Data Table 2 | Sequence identities between subunits of FluPol from C/Johannesburg/1/1966 and those from A/Little yellow-shouldered bat/Guatemala/060/2010 or B/Memphis/13/2003, calculated using EMBOSS Stretcher[45]

FluPol Subunit	Sequence Identity with C (%)	
	A	B
PB1	38.4	40.8
PB2	23.3	25.2
P3 / PA	25.4	25.6

Extended Data Table 3 | Domain position differences between apo-FluPolC and promoter-bound FluPolA, calculated using SHP

Domain	Rotation (°)	Distance (Å)
PA_{Endo} / $P3_{Endo}$	140	19
$PB2_{Mid}$ and $PB2_{Cap\text{-}627\ Linker}$	141	29
$PB2_{Cap}$	122	30
$PB2_{627}$	163	79
$PB2_{NLS}$	134	93

LETTER

doi:10.1038/nature15713

Structural insight into substrate preference for TET-mediated oxidation

Lulu Hu[1,2,3]*, Junyan Lu[4]*, Jingdong Cheng[1,2]*, Qinhui Rao[1,2], Ze Li[1,2], Haifeng Hou[5], Zhiyong Lou[6,7], Lei Zhang[1,2], Wei Li[1], Wei Gong[1,2], Mengjie Liu[1,2], Chang Sun[1,2], Xiaotong Yin[1,2], Jie Li[1,2], Xiangshi Tan[1], Pengcheng Wang[8], Yinsheng Wang[8], Dong Fang[9], Qiang Cui[9], Pengyuan Yang[1,2], Chuan He[10,11], Hualiang Jiang[4], Cheng Luo[4] & Yanhui Xu[1,2,3]

DNA methylation is an important epigenetic modification[1–3]. Ten-eleven translocation (TET) proteins are involved in DNA demethylation through iteratively oxidizing 5-methylcytosine (5mC) into 5-hydroxymethylcytosine (5hmC), 5-formylcytosine (5fC) and 5-carboxylcytosine (5caC)[4–8]. Here we show that human TET1 and TET2 are more active on 5mC-DNA than 5hmC/5fC-DNA substrates. We determine the crystal structures of TET2–5hmC-DNA and TET2–5fC-DNA complexes at 1.80 Å and 1.97 Å resolution, respectively. The cytosine portion of 5hmC/5fC is specifically recognized by TET2 in a manner similar to that of 5mC in the TET2–5mC-DNA structure[9], and the pyrimidine base of 5mC/5hmC/5fC adopts an almost identical conformation within the catalytic cavity. However, the hydroxyl group of 5hmC and carbonyl group of 5fC face towards the opposite direction because the hydroxymethyl group of 5hmC and formyl group of 5fC adopt restrained conformations through forming hydrogen bonds with the 1-carboxylate of NOG and N4 exocyclic nitrogen of cytosine, respectively. Biochemical analyses indicate that the substrate preference of TET2 results from the different efficiencies of hydrogen abstraction in TET2-mediated oxidation. The restrained conformation of 5hmC and 5fC within the catalytic cavity may prevent their abstractable hydrogen(s) adopting a favourable orientation for hydrogen abstraction and thus result in low catalytic efficiency. Our studies demonstrate that the substrate preference of TET2 results from the intrinsic value of its substrates at their 5mC derivative groups and suggest that 5hmC is relatively stable and less prone to further oxidation by TET proteins. Therefore, TET proteins are evolutionarily tuned to be less reactive towards 5hmC and facilitate the generation of 5hmC as a potentially stable mark for regulatory functions.

Previous studies have shown that 5hmC is 10- to 100-fold more abundant than 5fC/5caC and that its level is relatively high in neurons, self-renewing and pluripotent stem cells, and greatly reduced in cancer cells[6,7,10–13]. The depletion of *Tdg* in mouse embryonic stem cells leads to an accumulation of 5fC and 5caC by two- to ten fold, but no apparent changes in 5hmC and 5mC levels, suggesting that thymine-DNA glycosylase is not predominately responsible for the different abundance of 5hmC and 5fC/5caC[14]. *In vitro* biochemical analyses indicate that mouse Tet2 and *Naegleria* Tet-like protein possess higher activity for 5mC than for 5hmC/5fC[6,15], suggesting that TET enzymes might play a major role in controlling the level of 5mC-oxidized derivatives.

To understand how TET proteins recognize 5hmC/5fC and iteratively oxidize 5mC and its derivatives, we performed an *in vitro* enzymatic activity assay using purified recombinant catalytic domain of human TET1 or TET2 with the products detected by liquid-chromatography–tandem mass spectrometry (LC–MS/MS) (Extended Data Fig. 1 and Supplementary Tables 1 and 2). TET1 (12.5 μM) converted 89% 5mC to 47% 5hmC, 19% 5fC and 23% 5caC for 5mC-DNA substrate (Fig. 1a). In contrast, it could only oxidize 25% 5hmC or 18% 5fC. TET2 (5 μM) oxidized almost all 5mC-DNA, but 52% 5hmC-DNA or 33% 5fC-DNA (Fig. 1b). In low protein concentration (1 μM), TET2 could still oxidize over 90% 5mC, but only 15% 5hmC or negligible 5fC (Fig. 1c). Thus, human TET1 and TET2 both showed higher activity on 5mC-DNA than on 5hmC/5fC-DNA substrates, which is consistent with previous observations and suggests a conserved mechanism for TET proteins[6,15].

We next detected product generation at different time points (Fig. 1d and Supplementary Table 3). For 5mC-DNA substrate, TET2 (1 μM) converted 73% 5mC to 70% 5hmC and less than 3% 5fC at 5 min. A noticeable amount of 5fC only emerged at 10 min when 5hmC relatively accumulated (10.5 μM, ~68% of 5mC substrate), and detectable 5caC emerged at 20 min when 5fC accumulated (3.5 μM, ~23% of 5mC substrate). For 5hmC-DNA substrate, the level of 5fC was low at 5 min after initiation of the oxidation, whereas 5caC only emerged at 20 min when 5fC accumulated (1.2 μM, ~10% of 5hmC substrate) (Fig. 1e). No 5caC was detected when 5fC-DNA was used as the substrate (Fig. 1f). The results indicate that 5hmC was the major product of TET-mediated 5mC oxidation under our experimental conditions and considerable amounts of 5fC/5caC are not generated until 5hmC accumulates, which is consistent with the observation that cellular 5hmC is relatively stable and significantly more prevalent than 5fC/5caC[6,14].

We next performed steady-state kinetic analyses for the TET2-mediated oxidation of 5mC/5hmC/5fC-DNA substrates (Fig. 1g–i and Supplementary Table 4). We optimized TET2 concentration to ensure only one product was generated for each measurement (Extended Data Fig. 1c–h). For example, 0.5 μM TET2 converted 5mC-DNA to 5hmC-DNA, but not 5fC/5caC-DNA, under the experimental conditions. The kinetic analyses indicate that TET2 has higher activity for 5mC-DNA ($K_{cat}/K_m = 4.42 \times 10^3\,M^{-1}\,s^{-1}$) than for 5hmC-DNA ($K_{cat}/K_m = 0.70 \times 10^3\,M^{-1}\,s^{-1}$) or 5fC-DNA ($K_{cat}/K_m = 0.35 \times 10^3\,M^{-1}\,s^{-1}$).

A similar substrate preference was observed for 5mC/5hmC/5fC-DNA substrates with different sequences (AT-rich/CG-rich) and lengths (26/58/100 base pairs (bp)) (Extended Data Fig. 2a–e and Supplementary Table 5). The presence of CpG-DNA or different products did not significantly inhibit TET2 activity, indicating that the substrate preference does not result from product inhibition

[1]Fudan University Shanghai Cancer Center, Institute of Biomedical Sciences, Shanghai Medical College of Fudan University, Shanghai 200032, China. [2]Key Laboratory of Molecular Medicine, Ministry of Education, Department of Systems Biology for Medicine, School of Basic Medical Sciences, Shanghai Medical College of Fudan University, Shanghai 200032, China. [3]State Key Laboratory of Genetic Engineering, Collaborative Innovation Center of Genetics and Development, School of Life Sciences, Fudan University, Shanghai 200433, China. [4]Drug Discovery and Design Center, State Key Laboratory of Drug Research, Shanghai Institute of Materia Medica, Chinese Academy of Sciences, Shanghai 201203, China. [5]Beijing Synchrotron Radiation Facility, Institute of High Energy Physics, Chinese Academy of Sciences, Beijing 100049, China. [6]Laboratory of Structural Biology, Tsinghua University, Beijing 100084, China. [7]MOE Laboratory of Protein Science, School of Medicine, Tsinghua University, Beijing 100084, China. [8]Department of Chemistry, University of California-Riverside, Riverside, California 92521-0403, USA. [9]Theoretical Chemistry Institute, Department of Chemistry, University of Wisconsin-Madison, 1101 University Avenue, Madison, Wisconsin 53706, USA. [10]Department of Chemistry and Institute for Biophysical Dynamics, The University of Chicago, 929 East 57th Street, Chicago, Illinois 60637, USA. [11]Howard Hughes Medical Institute, The University of Chicago, 929 East 57th Street, Chicago, Illinois 60637, USA.
*These authors contributed equally to this work.

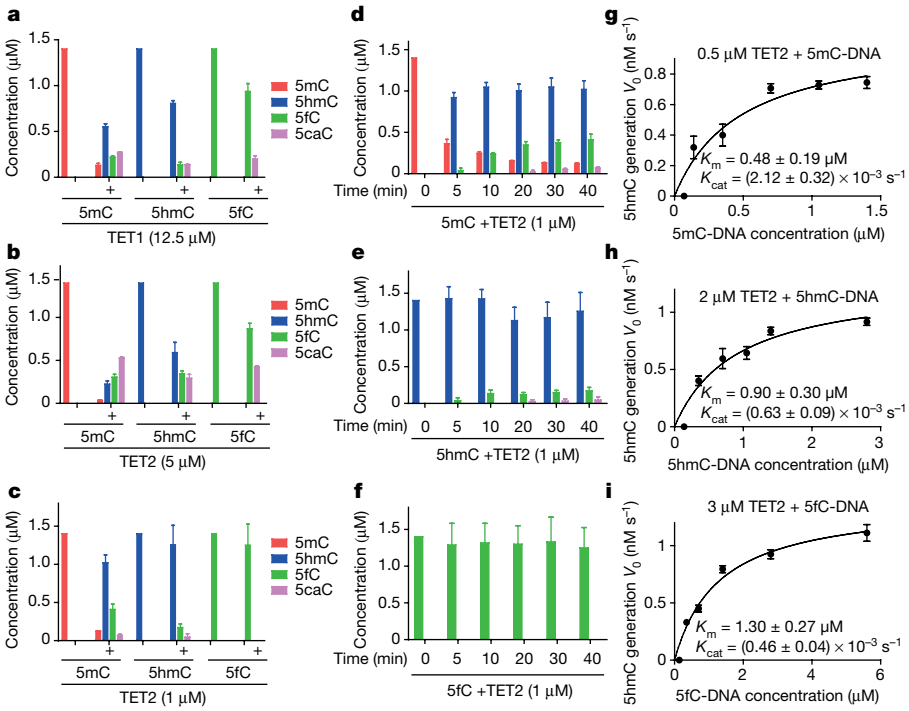


Figure 1 | TET proteins prefer 5mC-DNA, but not 5hmC/5fC-DNA as substrate. **a**–**c**, LC–MS/MS analyses of nucleoside hydrolytes for enzymatic assays using various DNA substrates and purified TET1 (**a**) or TET2 in two different concentrations (**b**, **c**). **d**–**f**, Enzymatic activities of TET2 on 5mC-DNA (**d**), 5hmC-DNA (**e**) or 5fC-DNA (**f**) at different time points. **g**–**i**, Michaelis–Menten plots of the steady-state kinetics for TET2-mediated oxidation of 5mC/5hmC/5fC-DNA. To match the linear interval in the first 2.5 min and generate only one product for all the reactions, 0.5 μM, 2 μM and 3 μM TET2 were used for the three measurements with 5mC-, 5hmC- and 5fC-DNA, respectively. The oxidized products were analysed by LC–MS/MS. Relative amounts of 5mC, 5hmC, 5fC and 5caC were calculated for each measurement according to standard curves of various cytosine derivatives. Error bars, s.d. for triplicate/duplicate experiments from three/two independent assays for **a**–**f** and **g**–**i**, respectively.

(Extended Data Fig. 2f–h and Supplementary Table 6). Taken together, TET1/2 has higher enzymatic activity for 5mC-DNA than for 5hmC/5fC-DNA substrates, suggesting an intrinsic property and conserved mechanism for TET-mediated oxidation.

To investigate the mechanism for substrate preference of TET proteins, we first measured the DNA-binding affinities, which may affect activity of TET2 on various substrates. Fluorescence polarization measurements indicated that TET2 has comparable DNA-binding affinity to C/5mC/5hmC/5fC-DNA (Fig. 2a). The presence of Fe^{2+}/Mn^{2+} and NOG/2-OG did not affect the DNA-binding affinity (Extended Data Fig. 3 and Extended Data Table 1a). Surface plasmon resonance (SPR) measurements showed no significant difference in the association or dissociation constant for the dynamic interaction between TET2 and 5mC/5hmC/5fC-DNA substrates (Fig. 2b–d and Extended Data Table 1b). Thus, TET2 binds to 5mC/5hmC/5fC-DNA with comparable DNA-binding affinity, which is unlikely to result in a substrate preference of TET2.

To investigate whether 5hmC/5fC adopts a non-optional conformation or multiple conformations (some catalytically productive and others not) within the catalytic cavity to hamper TET2-mediated oxidation, we determined the crystal structures of TET2–5hmC-DNA and TET2–5fC-DNA at 1.80 Å and 1.97 Å resolution, respectively (Fig. 3a and Extended Data Table 2). The two complexes adopt similar overall fold to that of TET2–5mC-DNA (Protein Data Bank (PDB) accession number 4NM6)[9] (Extended Data Fig. 4). Briefly, the Cys-rich domain and double-stranded β-helix (DSBH) domain together form a compact globular fold, with the catalytic DSBH core located in the centre and two highly conserved loops stabilizing the DNA right above the DSBH core (Fig. 3b).

TET2 binds to 5hmC-DNA and 5fC-DNA through extensive hydrogen bonds and hydrophobic interactions (Fig. 3c and Extended Data Fig. 5). TET2 recognizes 5hmC-DNA in a manner similar to that of 5mC-DNA in TET2–5mC-DNA complex. Higher-resolution and clearer electron density maps provide additional insight into the mechanism for substrate recognition by TET2. The guanine:hydroxymethylcytosine (G7:hmC7′) base pair of DNA forms a base-stacking interaction with residue Y1294 of TET2, and the hydroxymethyl group of hmC7′ is not recognized by TET2. In addition, the endocyclic oxygen atom O2 of hmC7′ forms water-mediated hydrogen bonds with residues Y1295 and R1302 of TET2 (Extended Data Fig. 5). In the 5fC-DNA structure, because hemi-formylated double-stranded DNA (dsDNA) was used for crystallization, it is C7′ that pairs with G7 of the CpG dinucleotide outside the catalytic cavity. Nevertheless, the G7:C7′ base pair of 5fC-DNA is stabilized by TET2 in a similar fashion to that observed in TET2–5mC/5hmC-DNA complex structures (Extended Data Fig. 5).

Except for CpG dinucleotide, no base of DNA is specifically recognized by TET2, suggesting that TET2 has no preference for DNA sequence apart from the CpG dinucleotide (Extended Data Fig. 5). Consistently, TET2 shows comparable enzymatic activity on DNA containing AT- or CG-rich sequences flanking the methyl-CpG dinucleotide (Extended Data Fig. 2 and Supplementary Table 5). The result agrees well with the genome-wide analyses, in which 5hmC/5fC/5caC occurs mainly on the CpG site but has no preference on its flanking sequence[14,16,17].

The hydroxymethylcytosine or the formylcytosine is flipped out of the DNA double helix and inserted into the catalytic cavity. As observed in the TET2–5mC-DNA structure, the cytosine potion of 5hmC/5fC forms hydrogen bonds with residues H1904 and N1387 of TET2, and the interaction is further supported by base-stacking interaction between residue Y1902 and the pyrimidine base of 5hmC/5fC (Fig. 3d–h and Extended Data Fig. 4e–g). Additional hydrogen bonds were observed in the TET2–5hmC-DNA complex, including one between residue H1386 and endocyclic oxygen atom O2 of base 5hmC, and a water-mediated hydrogen bond between exocyclic amino (N4) nitrogen and residue T1393 of TET2 (Fig. 3d). In the TET2–5fC-DNA

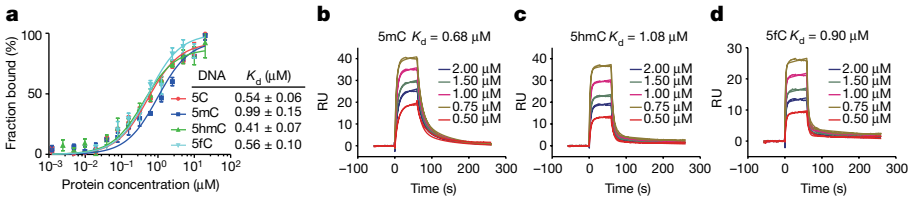


Figure 2 | DNA-binding affinity of TET2. a, The DNA-binding affinity of TET2 measured by fluorescence polarization. Fluorescein amidite (FAM)-labelled 18-bp C/5mC/5hmC/5fC-DNA (5 nM) was incubated with increasing amounts of TET2. Error bars, s.d. for triplicate experiments from three independent experiments. **b**–**d**, SPR measurements of the interaction between TET2 and biotinylated 26-bp DNA. The data were fitted with a two-state binding model with binding affinity (K_d) indicated. RU, resonance units.

complex, residue H1386 flips away from 5fC and has no direct contact with DNA substrate. A relative weaker water-mediated hydrogen bond (3.27 Å) is formed between base 5fC and residue T1393 of TET2 (Fig. 3e).

The pyrimidine bases of 5mC/5hmC/5fC adopt almost identical conformation within the catalytic cavity in the three compared complexes (Fig. 3f–h). The network of TET2–DNA interactions (within and outside the catalytic cavity) collectively offers the specific recognition of 5mC/5hmC/5fC-pG dinucleotide by TET2. Notably, the cytosine portion of 5hmC/5fC within the catalytic cavity adopts an almost identical conformation to that of 5mC in the TET2–5mC-DNA structure (Extended Data Fig. 4d), suggesting that substrate recognition is unlikely to result in substrate preference of TET2.

Structural comparison of 5mC/5hmC/5fC-DNA-TET2 indicates that the major difference exists in their 5mC derivative groups. Notably, the hydrophobic methyl group of 5mC directly points to the catalytic centre and has no contact with NOG or TET2 residues (Fig. 3f). In contrast, the hydroxymethyl group of 5hmC adopts restrained conformation through forming a hydrogen bond (~2.6 Å) with 1-carboxylate of NOG (Fig. 3g), whereas the formyl group of 5fC is restrained by an intramolecular hydrogen bond formed between the carbonyl group and the N4 exocyclic nitrogen of cytosine (Fig. 3h). As a result, the hydroxyl group of 5hmC and the carbonyl group of 5fC face towards opposite direction in the catalytic cavity when the two structures are superimposed. The conformational difference of 5mC/5hmC/5fC in these pre-catalysis complexes results from the intrinsic properties of their 5mC derivative groups. The above analyses suggest that such different intrinsic properties may also lead to distinct behaviour of 5mC/5hmC/5fC during the catalysis, and thus result in distinct efficiency for TET2-mediated oxidation.

Previous studies have proposed a consensus mechanism for 2-OG/Fe(II)-dependent dioxygenases, which mainly involves four steps of reaction (Fig. 4a). It has been proposed that hydrogen abstraction is the rate-controlling step for oxidation mediated by AlkB[18], which is

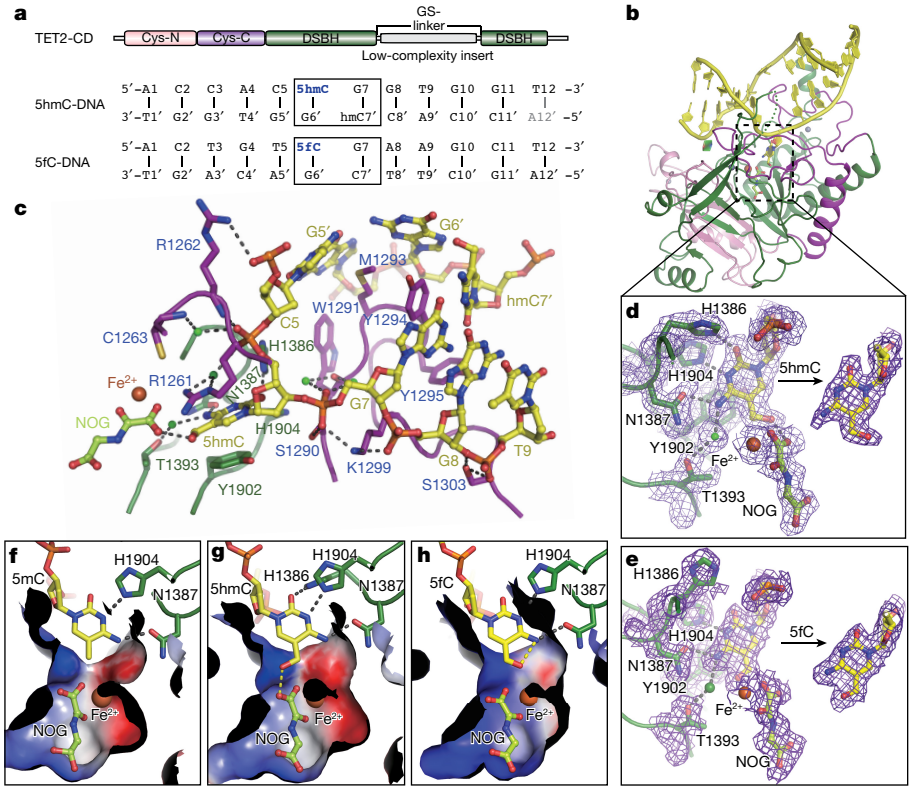


Figure 3 | Structure of TET2–5hmC-DNA complex. a, Colour-coded domain structure of the human TET2 catalytic domain and the sequences of 12-bp fully hydroxymethylated-DNA and hemi-formylated-DNA for crystallization. **b**, Ribbon representation of TET2–5hmC-DNA. NOG and the bases of DNA are shown in stick representations. An iron and three zinc cations are shown as red and grey balls, respectively. **c**, Interactions between TET2 and 5hmC-DNA with critical bases of DNA and residues of TET2 are shown in stick representations. **d**, **e**, Close-up views for the recognition of 5hmC (**d**) and 5fC (**e**) by TET2. The $2F_{observed} - F_{calculated}$ simulated annealing omit maps for residues involving 5hmC or 5fC recognition are shown. The omit maps for 5hmC and 5fC are indicated for clarity. The maps were calculated at 1.8 Å (TET2–5hmC-DNA) and 1.97 Å (TET2–5fC-DNA) respectively, and contoured at 1.0σ. Note that all critical groups, including the hydroxyl group of 5hmC and carbonyl group of 5fC, are well covered by the map, indicating that the structure models were built correctly. **f**–**h**, Close-up views for the recognition of 5mC (**f**, PDB 4NM6), 5hmC (**g**) and 5fC (**h**) by the catalytic cavity of TET2, which is shown in surface representation. The hydrogen bonds are indicated as dashed lines.

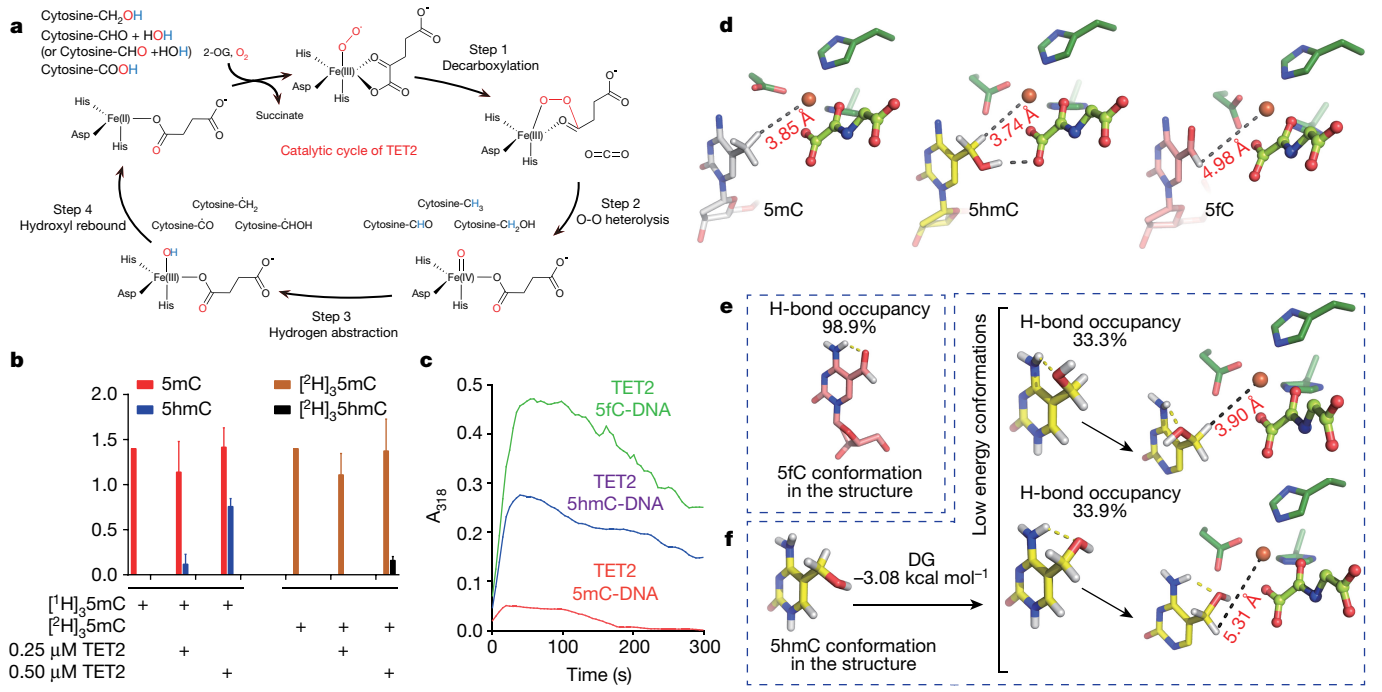


Figure 4 | Mechanism for substrate preference for TET-mediated oxidation. a, Model for the oxidative reactions catalysed by TET proteins. **b**, Decreased enzymatic activities of TET2 by substrate deuteration. The assays were performed as in Fig. 1 using 5mC-DNA ($[^1H]_3$5mC-DNA) and 5mC-DNA ($[^2H]_3$5mC-DNA) as substrates. To avoid ^2H exchange in aqueous phase, the reactions were performed using minimized enzyme and reaction times to prevent 5fC/5caC generation. Error bars, s.d. for triplicate experiments from three independent assays. **c**, The comparative kinetic traces from the reactions of TET2·Fe(II)·2-OG in the presence of 5mC/5hmC/5fC-DNA. The formation and decay of catalytic intermediate for each reaction was monitored by stopped-flow absorption at 318 nm. **d**, Hydrogens of 5mC derivatives are indicated in the structures of TET2–5mC/5fC-DNA complexes (pre-catalysis form). The closest distance between the abstractable hydrogen and Fe(II) is indicated as a dashed line. **e**, The occupancy (percentage) of the hydrogen bond during 10 ns molecular dynamic simulation of free 5fC. The intramolecular hydrogen bond formed in 5fC is indicated as a yellow dashed line. **f**, Comparison of the free energies of 5hmC in different conformations suggests that 5hmC also has the tendency to form an intramolecular hydrogen bond. The occupancy (percentage) of the intramolecular hydrogen bond is shown as in **e**. Note that the conformation of one 5hmC (bottom) is similar to that of 5fC in TET2–5fC-DNA, in which the abstractable hydrogen is away from the Fe(II) if the low energy conformation of 5hmC is positioned into the active site.

structurally and mechanically similar to TET proteins[9,19,20]. To test whether hydrogen abstraction is the rate-controlling step of TET2-mediated substrate oxidation, we measured the enzymatic activity of TET2 using regular 5mC-DNA ($[^1H]_3$5mC-DNA) and deuterated-5mC-DNA ($[^2H]_3$5mC-DNA) in which all of the hydrogen atoms of the methyl group were replaced by deuterium. The introduction of deuterium at the reactive position of 5mC-DNA substrate leads to a significant decrease in the enzymatic activity (Fig. 4b and Supplementary Table 7), indicating an obvious kinetic isotope effect. Notably, 0.25 μM TET2 oxidized ~10% of $[^1H]_3$5mC-DNA into $[^1H]_3$5hmC-DNA but showed undetectable activity towards $[^2H]_3$5mC-DNA. When treated with 0.5 μM TET2, 54% of $[^1H]_3$5mC-DNA was converted into $[^1H]_3$5hmC-DNA whereas only 11% of $[^2H]_3$5mC-DNA was oxidized. The result is consistent with previous studies of taurine α-ketoglutarate dioxygenase (tauD)[21] and supports the idea that hydrogen abstraction is the key step for TET-mediated oxidation of 5mC.

In the hydrogen abstraction step, a ferryl-oxo (Fe(IV) = O) intermediate (positive feature at 318 nm) is formed upon decarboxylation of 2-OG and oxidation is initiated (decay of the ferryl-oxo intermediate) by abstraction of a hydrogen atom from the target carbon of the substrate[20,21] (Fig. 4a). To test whether hydrogen abstraction contributes to the substrate preference of TET2, we measured the formation and decay of the ferryl-oxo intermediate for 5mC/5hmC/5fC-DNA substrates using stopped-flow spectroscopy, assuming the 318 nm species represents a catalytically valid intermediate according to previous studies[21]. The result shows different kinetic processes for TET-mediated oxidation of 5mC/5hmC/5fC-DNA. Comparison of the 318 nm kinetic traces indicates that the amplitude is significantly greater (more ferryl-oxo intermediate accumulation) for the reactions using 5hmC/5fC-DNA than for those using 5mC-DNA (Fig. 4c and Extended Data Fig. 6). This feature persisted much longer (slower decay of the ferryl-oxo intermediate) for the reactions using 5hmC/5fC-DNA than those using 5mC-DNA. For all the three reactions, decay but not formation of the ferryl-oxo intermediate takes much longer, supporting the hypothesis that the hydrogen abstraction after formation of ferryl-oxo intermediate accounts for the substrate preference of TET2.

Previous experimental and computational studies have suggested that higher homolytic C–H bond dissociation energy (BDE) of substrates would lead to lower hydrogen abstraction efficiency[22,23]. We therefore performed the calculations and found that the C–H BDEs for the 5-substitution group of 5mC, 5hmC and 5fC do not strictly follow the order of abstraction efficiencies (Fig. 4c and Extended Data Table 3). The C–H BDE for the formyl group of 5fC is slightly higher (~1 kcal mol^{-1}) than that of 5mC, but the C–H BDE for the hydroxymethyl group of 5hmC is the lowest, suggesting that other factors may influence the abstraction efficiency.

Structural analyses indicate that the abstractable hydrogen of 5fC is relatively far away (4.98 Å) from the iron because of its intramolecular hydrogen bond[24] (Figs 3h and 4d). Such planar conformation may prevent the abstractable hydrogen adopting a favourable orientation for abstraction reaction and thus result in low catalytic efficiency. In contrast, the abstractable hydrogens in 5mC have no such restriction (C–C bond can freely rotate) and would adopt a favourable conformation for hydrogen abstraction. As for the 5hmC in the pre-catalysis structure, the hydroxyl group forms a hydrogen bond with the C-1 carboxyl group of 2-OG, which positions one of its abstractable hydrogens close (3.74 Å) to the iron (Fig. 4d). However, this hydrogen bond would not be maintained after the decarboxylation of 2-OG, in which the C-1 carboxyl group is converted into a CO_2 molecule

(Fig. 4a). Further calculations indicate that free 5hmC and 5fC have the tendency to form intracellular hydrogen bonds, which may prevent the hydrogen abstraction and lead to the reduced activity (Fig. 4e, f). Crystal structures of the catalytic intermediate states of TET2 would provide additional structural evidence for understanding the mechanism for TET-mediated oxidation. Note that hydrogen abstraction is not always the rate-limiting step for 2-OG/Fe(II)-dependent dioxygenases. Interestingly, iterative oxidation on different substrates was observed in several 2-OG/Fe(II)-dependent dioxygenases[20,25–27]. Thus, it is of interest to test whether a similar mechanism applies for iterative oxidation mediated by other 2-OG/Fe(II)-dependent dioxygenases.

In summary, this work reveals that TET proteins are more active on 5mC-DNA than 5hmC/5fC-DNA substrates, which results from the distinct intrinsic properties of 5mC/5hmC/5fC within the catalytic cavity of TET proteins during oxidation. Thus, once established in the genome, 5hmC is less prone to further oxidization, unless TET proteins are stimulated to be more active. Regulation of the genomic localization and/or enzymatic activity of TET proteins may precisely control the patterns of 5mC and its oxidized derivatives. For example, vitamin C enhances TET activity and significantly increases levels of 5hmC/5fC/5caC in mouse embryonic stem cells and regulates somatic cell reprogramming[28]. TET proteins are therefore evolutionarily tuned to be less reactive towards 5hmC, perhaps to facilitate its generation as a potentially stable mark for regulatory functions. Genome-wide analyses indicate that 5fC and 5caC are mainly observed in specific genomic regions[14,16], suggesting that TET might be more concentrated or active in these regions by its interacting proteins or some mechanisms yet to be discovered. It will also be of interest to investigate the mechanism by which TET proteins are either activated to generate more 5fC/5caC for DNA demethylation or to retain relatively low activity to generate 5hmC *in vivo*.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 9 May; accepted 10 September 2015.
Published online 28 October 2015.

1. Bird, A. DNA methylation patterns and epigenetic memory. *Genes Dev.* **16,** 6–21 (2002).
2. Jaenisch, R. & Bird, A. Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nature Genet.* **33** (Suppl.), 245–254 (2003).
3. Smith, Z. D. & Meissner, A. DNA methylation: roles in mammalian development. *Nature Rev. Genet.* **14,** 204–220 (2013).
4. Guo, F. *et al.* Active and passive demethylation of male and female pronuclear DNA in the mammalian zygote. *Cell Stem Cell* **15,** 447–458 (2014).
5. He, Y. F. *et al.* Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA. *Science* **333,** 1303–1307 (2011).
6. Ito, S. *et al.* Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. *Science* **333,** 1300–1303 (2011).
7. Tahiliani, M. *et al.* Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science* **324,** 930–935 (2009).
8. Wang, L. *et al.* Programming and inheritance of parental DNA methylomes in mammals. *Cell* **157,** 979–991 (2014).
9. Hu, L. *et al.* Crystal structure of TET2-DNA complex: insight into TET-mediated 5mC oxidation. *Cell* **155,** 1545–1555 (2013).
10. Globisch, D. *et al.* Tissue distribution of 5-hydroxymethylcytosine and search for active demethylation intermediates. *PLoS One* **5,** e15367 (2010).
11. Mellén, M., Ayata, P., Dewell, S., Kriaucionis, S. & Heintz, N. MeCP2 binds to 5hmC enriched within active genes and accessible chromatin in the nervous system. *Cell* **151,** 1417–1430 (2012).
12. Pfaffeneder, T. *et al.* The discovery of 5-formylcytosine in embryonic stem cell DNA. *Angew. Chem. Int. Edn Engl.* **50,** 7008–7012 (2011).
13. Song, C. X. *et al.* Selective chemical labeling reveals the genome-wide distribution of 5-hydroxymethylcytosine. *Nature Biotechnol.* **29,** 68–72 (2011).
14. Shen, L. *et al.* Genome-wide analysis reveals TET- and TDG-dependent 5-methylcytosine oxidation dynamics. *Cell* **153,** 692–706 (2013).
15. Hashimoto, H. *et al.* Structure of a *Naegleria* Tet-like dioxygenase in complex with 5-methylcytosine DNA. *Nature* **506,** 391–395 (2014).
16. Song, C. X. *et al.* Genome-wide profiling of 5-formylcytosine reveals its roles in epigenetic priming. *Cell* **153,** 678–691 (2013).
17. Yu, M. *et al.* Base-resolution analysis of 5-hydroxymethylcytosine in the mammalian genome. *Cell* **149,** 1368–1380 (2012).
18. Liu, H., Llano, J. & Gauld, J. W. A DFT study of nucleobase dealkylation by the DNA repair enzyme AlkB. *J. Phys. Chem. B* **113,** 4887–4898 (2009).
19. Yang, C. G. *et al.* Crystal structures of DNA/RNA repair enzymes AlkB and ABH2 bound to dsDNA. *Nature* **452,** 961–965 (2008).
20. Aik, W., McDonough, M. A., Thalhammer, A., Chowdhury, R. & Schofield, C. J. Role of the jelly-roll fold in substrate binding by 2-oxoglutarate oxygenases. *Curr. Opin. Struct. Biol.* **22,** 691–700 (2012).
21. Price, J. C., Barr, E. W., Glass, T. E., Krebs, C. & Bollinger, J. M., Jr. Evidence for hydrogen abstraction from C1 of taurine by the high-spin Fe(IV) intermediate detected during oxygen activation by taurine:alpha-ketoglutarate dioxygenase (TauD). *J. Am. Chem. Soc.* **125,** 13008–13009 (2003).
22. Kaizer, J. *et al.* Nonheme FeIVO complexes that can oxidize the C–H bonds of cyclohexane at room temperature. *J. Am. Chem. Soc.* **126,** 472–473 (2004).
23. de Visser, S. P., Kumar, D., Cohen, S., Shacham, R. & Shaik, S. A predictive pattern of computed barriers for C–H hydroxylation by compound I of cytochrome P450. *J. Am. Chem. Soc.* **126,** 8362–8363 (2004).
24. Münzel, M. *et al.* Improved synthesis and mutagenicity of oligonucleotides containing 5-hydroxymethylcytosine, 5-formylcytosine and 5-carboxylcytosine. *Chemistry* **17,** 13782–13788 (2011).
25. Wondrack, L. M., Hsu, C. A. & Abbott, M. T. Thymine 7-hydroxylase and pyrimidine deoxyribonucleoside 2′ -hydroxylase activities in *Rhodotorula glutinis*. *J. Biol. Chem.* **253,** 6511–6515 (1978).
26. Baggaley, K. H., Brown, A. G. & Schofield, C. J. Chemistry and biosynthesis of clavulanic acid and other clavams. *Nat. Prod. Rep.* **14,** 309–333 (1997).
27. Fu, Y. *et al.* FTO-mediated formation of N^6-hydroxymethyladenosine and N^6-formyladenosine in mammalian RNA. *Nature Commun.* **4,** 1798 (2013).
28. Blaschke, K. *et al.* Vitamin C induces Tet-dependent DNA demethylation and a blastocyst-like state in ES cells. *Nature* **500,** 222–226 (2013).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank the staff of beamlines BL17U and BL19U at Shanghai Synchrotron Radiation Facility for assistance in data collection, and staff of the Biomedical Core Facility, Fudan University, for their help with mass spectrometry. The computation resources were supported by the Computer Network Information Center, Chinese Academy of Sciences and Shanghai Supercomputing Center. This work was supported by grants from the National Basic Research Program of China (2011CB965300), the National Science & Technology Major Project 'Key New Drug Creation and Manufacturing Program' of China (2014ZX09507-002), the National Natural Science Foundation of China (U1432242, 31425008, 91419301, 91313000, 31270779, 21210003), the Basic Research Project of Shanghai Science and Technology Commission (12JC1402700), the Program of Shanghai Subject Chief Scientist (14XD1400500), the Hi-Tech Research and Development Program of China (2012AA020302 and 2012AA01A305) and the Chinese Academy of Sciences (XDA01040305).

Author Contributions L.H., J.L., J.C., C.L. and Y.X. designed the experiments. L.H., J.C., Z.L., Q.R., W.G., M.L., C.S. and X.Y. purified protein. L.H. performed crystallization and all enzymatic assays. L.H., J.C., H.H., Z.L. and J.L. collected the data and determined the crystal structure. J.C. and L.H. performed the fluorescence polarization and SPR assays. L.H., J.C., W.L. and X.T. performed stopped-flow analyses. L.Z. and P.Y. designed and performed LC–MS/MS and analysed the data. J.L., H.J., C.L., D.F. and Q.C. designed and performed computational studies and analysed the data. C.H. and Y.W. provided standard for LC–MS/MS analyses. L.H., J.L., C.L. and Y.X. analysed the data and wrote the manuscript. Y.X. supervised the project.

Author Information The coordinates and structure factors for the TET2–5hmC-DNA and TET2–5fC-DNA structures have been deposited in the Protein Data Bank under accession numbers 5DEU and 5D9Y, respectively. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to Y.X. (xuyh@fudan.edu.cn) or C.L. (cluo@simm.ac.cn).

METHODS

No statistical methods were used to predetermine sample size. The experiments were not randomized. The investigators were not blinded to allocation during experiments and outcome assessment.

Protein expression and purification. The procedure for protein expression and purification of TET2 has been described previously[9]. In brief, the open reading frame corresponding to the catalytic domain of human TET1 (1418–2136) or TET2 (1129–1936 with residues 1481–1843 replaced by a 15-residue GS-linker GGGGSGGGGSGGGGS) was sub-cloned into modified pGEX-6p-1 or pET-28b vector and the plasmids were transformed into *Escherichia coli* strain BL21(DE3). The transformants were grown at 37 °C to an absorbance at 600 nm of 0.8 and induced by adding 0.1 mM isopropyl-β-D-thiogalactopyranoside. After further culture at 16 °C for 14–18 h, the cells expressing TET1 or TET2 were lysed and the supernatant was subjected to Ni-NTA columns for affinity purification. His/GST-tag was removed by on-column digestion at 4 °C for 12–16 h. The eluted proteins were further purified by ion exchange and gel filtration chromatography. The purified proteins were subjected to SDS–polyacrylamide gel electrophoresis, stained by Coomassie blue and visualized on a Tanon-5200 Chemiluminescent Imaging System (Tanon Science & Technology). The proteins were concentrated to 25 mg ml^{-1} and used for *in vitro* assays and crystallization.

TET2 enzymatic assays and LC–MS/MS analysis. Procedures for *in vitro* enzymatic assays and LC–MS/MS were described previously[9]. In brief, various DNA substrates was incubated with TET1 or TET2 in buffer containing 50 mM HEPES pH 8.0, 100 mM NaCl, 100 μM $Fe(NH_4)_2(SO_4)_2$, 2 mM ascorbate, 1 mM DTT and 1 mM ATP at 37 °C. Reactions were stopped by the addition of ten volumes of Buffer PN (Qiagen), and the DNA products were then purified using a QIAquick Nucleotide Removal Kit (Qiagen) following the manufacturer's instructions. The purified DNA products were denatured at 100 °C for 10 min and further digested to nucleosides with 0.5 U nuclease P1 (Sigma Aldrich) at 37 °C for 16 h and 0.5 U CIP (NEB) at 37 °C for 1.5 h. For the product inhibition assay, 1 μM TET2 was incubated with biotinylated 26-bp 5mC/5hmC/5fC-DNA substrates in the presence of corresponding biotin-free 5hmC/5fC/5caC-DNA with the same sequence. Unmodified CpG-containing DNA was used as control. For the deuterium isotope effect assay, 0.25 μM or 0.5 μM TET2 was incubated with biotinylated 20-bp $[^2H]_3$5mC-/$[^1H]_3$5mC-DNA substrates. After reaction for 10 min at 37 °C, followed by heat at 65 °C to deactivate TET2, the biotinylated 20-bp $[^1H]_3$5mC/$[^2H]_3$5mC-DNA was purified by strepavidin beads, and then treated as described above. The samples were subjected to LC–MS/MS using Shimadzu LC (LC-20AB pump) system. The amounts of 5mC derivatives were calculated according to the external standard curves for $[^1H]_3$5mC, $[^2H]_3$5mC, $[^1H]_3$5hmC, $[^2H]_3$5hmC, 5fC, 5caC and guanine (Extended Data Fig. 1).

Preparation of DNA substrates. All DNA duplexes (summarized in the table below) were synthesized by Generay Biotech and annealed from single-stranded primers. A palindromic 12-bp dsDNA (5′-ACCAC(C^{hm})GGTGGT-3′, C^{hm} = 5-hydroxymethyldeoxycytosine) and a hemi-formyl dsDNA (top strand, 5′-ACTGT(C^fG)AAGCT-3′; bottom strand, 5′-AGCTTCGACAGT-3′; C^f = 5-formyldeoxycytosine) were used for crystallization. Palindromic 12-bp dsDNAs (D1: top strand, 5′-ACCACXGGTGGT-3′; *X* = 5mC, 5hmC or 5fC) were used for stopped-flow spectrometry analyses. FAM-labelled palindromic 18-bp dsDNAs (D2: top strand, 5′-FAM-CAGCACACXGGTGTGCTG-3′; *X* = C, 5mC, 5hmC, 5fC or 5caC) were used for fluorescence polarization measurements. Biotinylated 26-bp dsDNAs (D3: top strand, 5′-biotin-CAGTAGTCTGGAC ACACXGGTCATGA-3′; bottom strand, 5′-TCATGACXGGTGTGTCCAGACTA CTG-3′; X = 5mC, 5hmC or 5fC) were used for SPR analyses. Palindromic 58-bp dsDNAs (D5: top strand, 5′-ACGATCAGATCCTAAGGCATCAGCACACXGGT GTGCTGATGCCTTAGGATCTGATCGT-3′; *X* = 5mC, 5hmC or 5fC) were used for enzymatic assays. To test the effect of flanking DNA sequence besides CpG dinucleotide on the substrate preference of TET2, we synthesized AT-rich dsDNAs (D6: top strand, 5′-ACCAGCAGATGGCCAGGCATCAGATATAXGTATATCTG ATGCCTGGCCATCTGCTGGT-3′; *X* = 5mC, 5hmC or 5fC) and CG-rich 58-bp dsDNAs (D7: top strand, 5′-ACTCAACAGACTACACAGTAGTGCCCCCXGCC CAGATGCTATTCAGTAACTGACACTG-3′; bottom strand, 5′-CAGTGTCAGT TACTGAATAGCATCTGGGXGGGGGGCACTACTGTGTAGTCTGTTGAGT-3′; *X* = 5mC, 5hmC or 5fC), compared with the 58-bp dsDNAs (D5). To test the effect of DNA length on the substrate preference of TET2, we synthesized 100-bp dsDNAs (D8: top strand, 5′-GCTTGGAGGTCCAAGCTAGCTACGATCAGATC CTAAGGCATCAGCACACXGGTGTGCTGATGCCTTAGGATCTGATCGTAG CTAGCTTGGACCTCCAAGC-3′; *X* = 5mC, 5hmC or 5fC), compared with 26-bp dsDNAs (D4: top strand, 5′-CAGTAGTCTGGACACACXGGTCATGA-3′; bottom strand, 5′-TCATGACXGGTGTGTCCAGACTACTG-3′; *X* = 5C, 5mC, 5hmC, 5fC or 5caC) and 58-bp dsDNAs (D5). To test product inhibition on TET2 activity, we used biotinylated 26-bp dsDNAs (D3) and corresponding biotin-free 26-bp dsDNAs (D4) with the same sequences. Biotinylated 20-bp dsDNA with deuterium-replaced 5mC (D9: top strand, 5′-biotin-CTTGGACACACXGGTCATGA-3′; bottom strand, 5′-TCATGACCmGGTGTGTCCAAG-3′; *X* = $[^2H]_3$5mC) or regular 5mC (D10: top strand, 5′-biotin-CTTGGACACACXGGTCATGA-3′; bottom strand, 5′-TCATGACCmGGTGTGTCCAAG-3′; *X* = 5mC) were used to test the isotope effect on TET2 activity. The DNAs used in this study are summarized in Supplementary Table 1.

Protein crystallization. Procedures for protein purification of TET2 have been described previously[9]. In brief, human TET1 (1418–2136) or TET2 (1129–1936 with residues 1481–1843 replaced by a 15-residue GS-linker) was purified to homogeneity for crystallization and enzymatic activity assays. For simplicity, the two proteins were designated as TET1 and TET2 in this work. Sequences of DNA used for crystallization and assays are described in the Preparation of DNA substrates section above. The crystals of human TET2 in complex with 12-bp 5hmC containing DNA were obtained using the hanging-drop, vapour-diffusion method by mixing 1 μl protein–DNA complex solution (25 mg ml^{-1}) with 1 μl reservoir solution containing 0.1 M MES (pH 6.4), 26% PEG monomethyl ether 2000 at 277 K. For crystallization of TET2 and hemi-formylated dsDNA complex, 1.5 μl protein–DNA complex solution (20 mg ml^{-1}) and 1.5 μl reservoir solution containing 0.1 M MES (pH 6.3), 21% PEG monomethyl ether 2000 were mixed and equilibrated by hanging-drop, vapour-diffusion at 277 K.

Data collection and structure determination. The data of TET2–5hmC-DNA and TET2–5fC-DNA were collected at wavelengths of 0.9792 Å and 0.97876 Å at Shanghai Synchrotron Radiation Facility beamlines BL17U and BL19U, respectively. Data were indexed, integrated and scaled using program HKL2000 (ref. 29). The structure was solved by molecular replacement using the TET2–5mC-DNA complex structure (PDB 4NM6) as the searching model[9]. The initial models were manually built with COOT[30] and refined using PHENIX package[31]. The quality of the final model was validated with the program MolProbity[32], indicating that 98.3% of residues were in favoured regions, 1.5% in allowed regions and 0.2% in outlier regions for TET2–5hmC-DNA, and 97.1% residues were in favoured regions, 2.7% in allowed regions and 0.2% in outlier regions for TET2–5fC-DNA. All structure figures were generated using PyMOL[33].

Fluorescence polarization measurements. Various modifications of FAM-18-bp DNA (5 nM) were mixed with increasing amounts of TET2. The mixtures were incubated in buffer containing 10 mM HEPES pH 7.0, 100 mM NaCl for 30 min at 4 °C. To measure DNA-binding affinity for different substrates under catalytic conditions, buffer A containing 100 μM Fe^{2+} and 1 mM NOG, and buffer B containing 100 μM Mn^{2+} and 1 mM 2-OG, were used to mimic catalytic conditions. In addition, buffer C containing 100 μM Fe^{2+} and 1 mM succinate was used to mimic the product release condition. Fluorescence polarization measurements were performed at 25 °C on a Synergy 4 Microplate Reader (BioTek). The data from three independent experiments were fitted using GraphPad Prism 5.

SPR measurements. Biotinylated-DNA was coupled to SA-chip (GE Healthcare) with a response of 20–30 response units, which was achieved by adjusting the concentration of the oligonucleotides and the time of contact. All SPR measurements were performed using a BIAcore T-100 instrument in running buffer containing 10 mM HEPES pH 7.4, 150 mM NaCl, 0.005% surface P20 at a flow rate of 30 μl min^{-1} and a temperature of 25 °C. Increasing concentrations of TET2 (0.5 μM, 0.75 μM, 1 μM, 1.5 μM, 2 μM) were injected into the same surface in running buffer for 60 s. The surface was washed with running buffer for 200 s after the dissociation of the complexes. The data were analysed by fitting all curves using a two-state binding model to determine the kinetics association and dissociation rate constants in Biacore T100 evaluation software.

Stopped-flow spectrometry. Stopped-flow kinetic experiments were performed at 25 °C with an SF-61 DX2 double-mixing instrument and a Xe lamp (SF-61DX2, TgK Scientific). For all experiments in this study, reactions were monitored in PM mode at 318 nm (to monitor the ferryl-oxo intermediate). Reaction mixture A containing 0.5 mM TET2 and 1 mM $Fe(NH_4)_2(SO_4)_2$ in O_2-free buffer solution (10 mM HEPES 7.0, 100 mM NaCl, 10 mM β-ME) was prepared under high pure N_2-atmosphere in an MBraun glove box. Reaction mixture B containing 2 mM 2-OG, 1 mM ATP and 0.5 mM 12-bp 5mC/5hmC/5fC dsDNA in the buffer solution (10 mM HEPES 7.0, 100 mM NaCl, 10 mM β-ME) was prepared in air environment. The two reactions (A and B) were mixed rapidly and the ferryl-oxo intermediate was monitored at 318 nm. Absorbance changes as a function of time were recorded and all curves were plotted with Origin software.

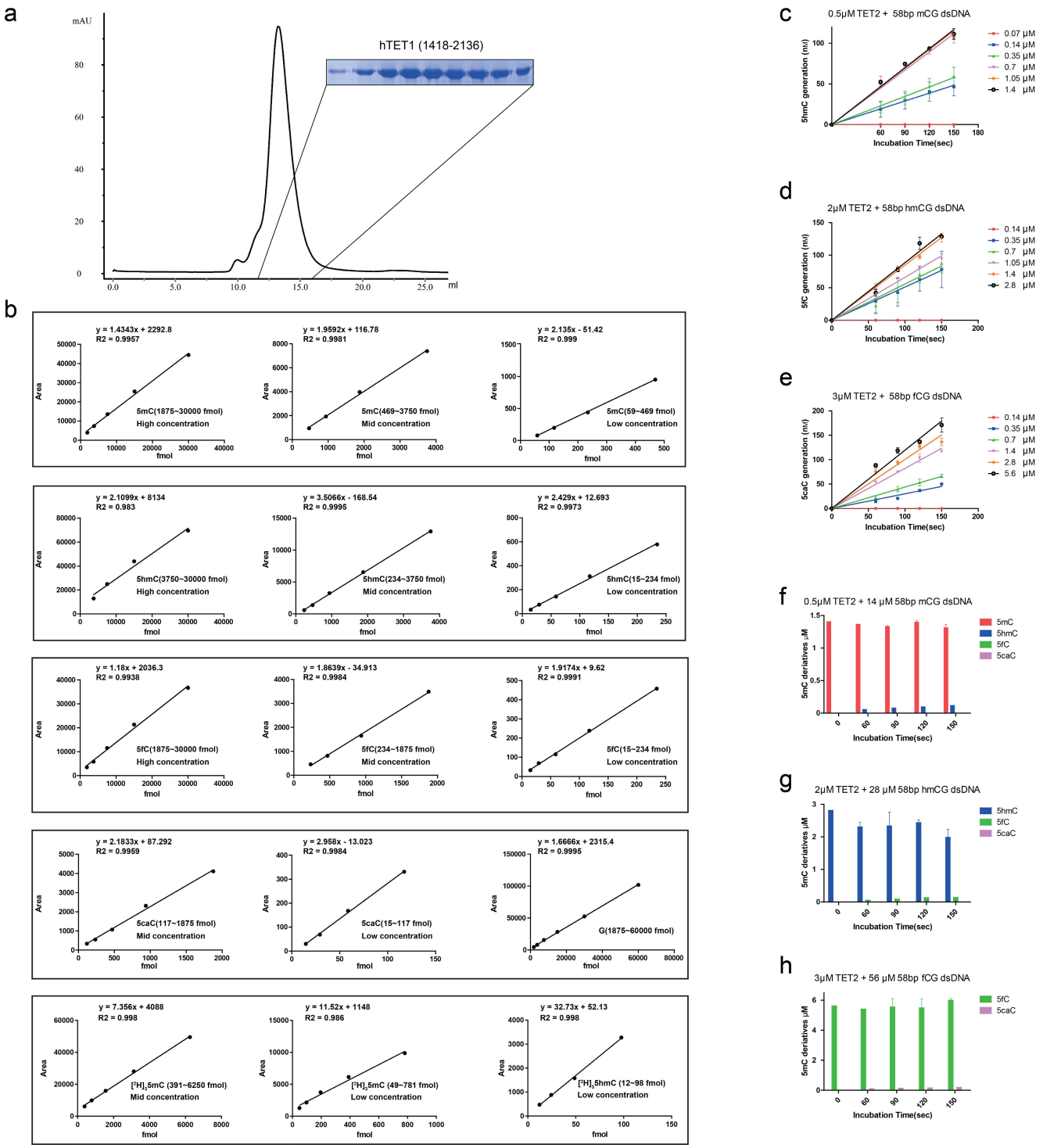
Computational details. The homolytic C–H BDEs for the 5-substitution groups of 5mC/5hmC/5fC were calculated using Gaussian 09 (ref. 34) as the enthalpy change of the following reaction at 298.15 K and 101.3 kPa: R–H → R• + H•, where R–H, R• and H• represent the parent base, the corresponding radical and the hydrogen atom, respectively. The initial geometry of each species was optimized at the UB3LYP/6-311+G(d,p) level. Frequency calculations at the same level were

conducted to verify that the optimized structures were the real minima without any imaginary vibration frequency. The single point energy and thermal corrections for the optimized structures were calculated by using several high-precision composite methods implemented in Gaussian 09, including CBS-QB3 (ref. 35), G4 (ref. 36), G3(MP2B3) (ref. 37) and CBS-QB3, with a conductor-like polarizable continuum model (C-PCM) implicit water model[38].

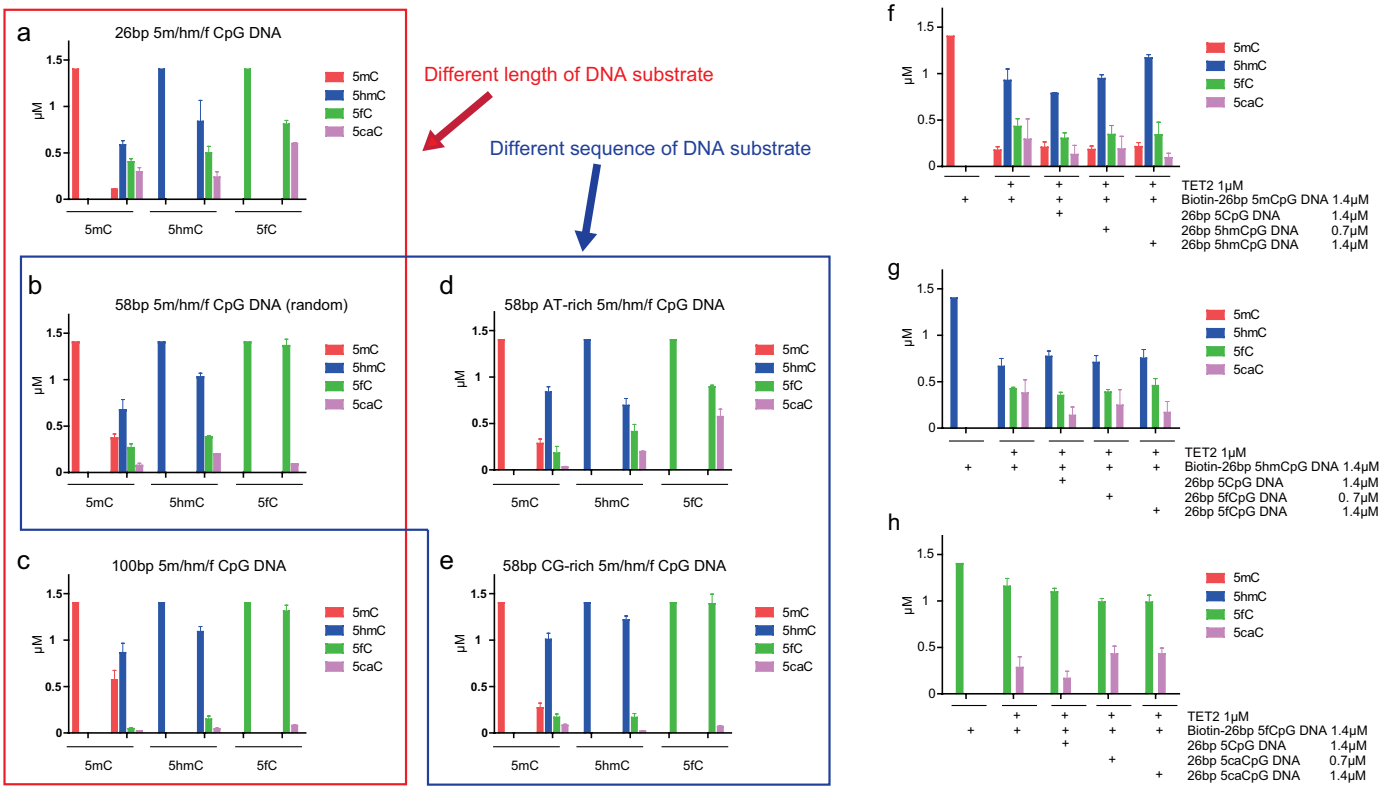
To identify the low energy conformations of 5hmC, a relaxed potential energy surface scan on the dihedral angle of the bond between the C5 carbon and the carbon atom of the hydroxymethyl group was performed at the B3LYP/6-31G(d,p) level. The conformations with the lowest energies on the potential energy surface were then fully optimized at the B3LYP/6-311+G(d,p) level. The free energies for the low energy conformations and the conformations observed in the TET2–5hmC crystal structure were calculated by using the CBS-QB3 method[35].

Molecular dynamics simulations of the free 5hmC and 5fC nucleotides were performed using the Amber 11 package[39]. The semi-empirical AM1 method was used to describe the nucleotide. The generalized-Born implicit solvent model[40] was used to mimic the environment within the binding pocket. Molecular dynamics simulation (10 ns) was performed for each model and 10,000 snapshots from the molecular dynamics trajectory were used to estimate the occupancy of the intramolecular hydrogen bond. The criteria for hydrogen bond formation were defined as (1) the O–H distance less than 2.5 Å and (2) the N–H–O angle larger than 90°.

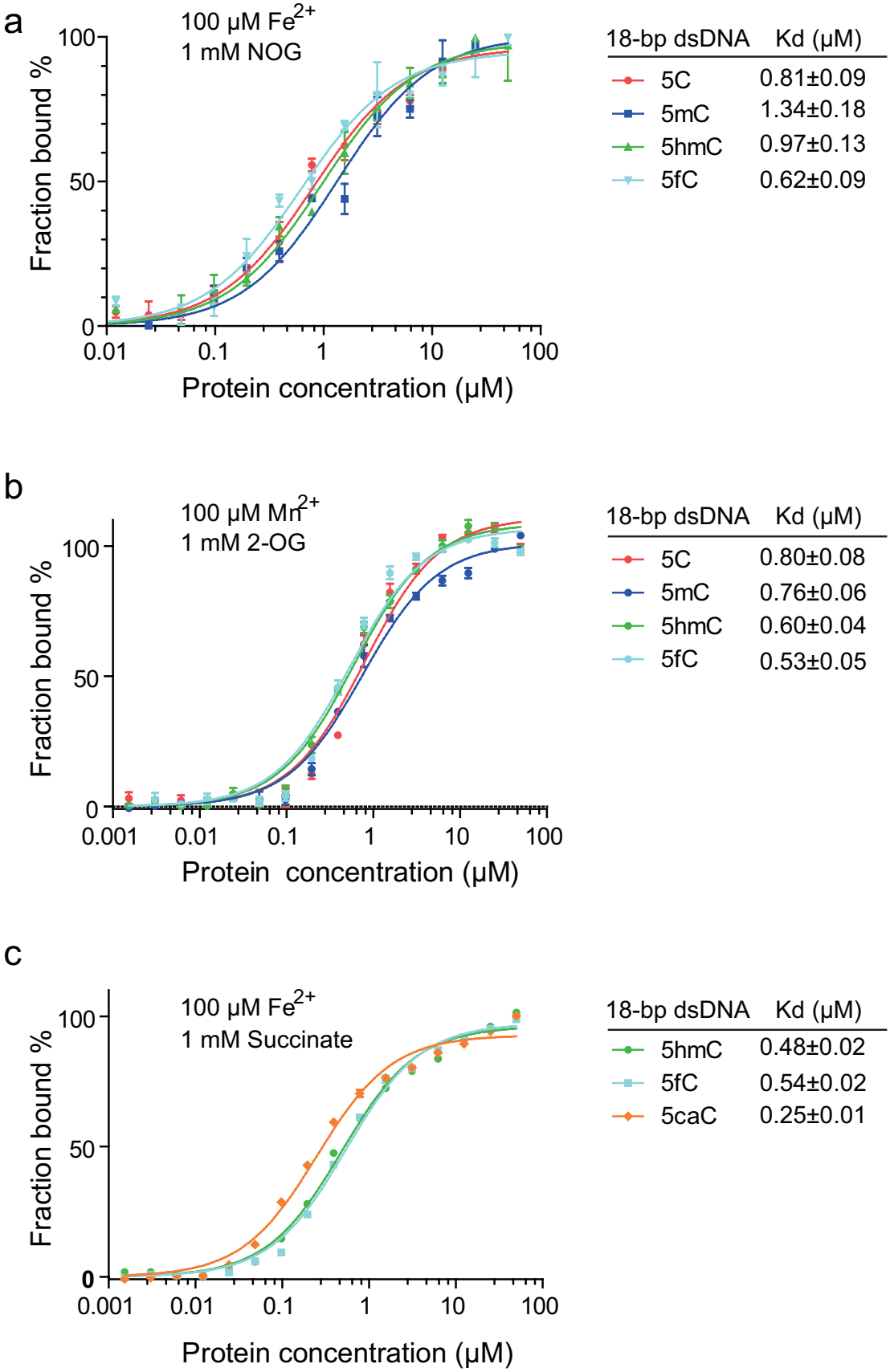
29. Otwinowski, Z. & Minor, W. Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol.* **276,** 307–326 (1997).
30. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr. D* **60,** 2126–2132 (2004).
31. Adams, P. D. *et al.* PHENIX: building new software for automated crystallographic structure determination. *Acta Crystallogr. D* **58**, 1948–1954 (2002).
32. Chen, V. B. *et al.* MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr. D* **66,** 12–21 (2010).
33. DeLano, W. L. The PyMOL molecular graphics system (DeLano Scientific, 2002).
34. Frisch, M. *et al. Gaussian 09, Revision A.02* (Gaussian, 2009).
35. Montgomery, J. A. Jr, Frisch, M. J., Ochterski, J. W. & Petersson, G. A. A complete basis set model chemistry. VII. Use of the minimum population localization method. *J. Chem. Phys.* **112,** 6532–6542 (2000).
36. Curtiss, L. A., Redfern, P. C. & Raghavachari, K. Gaussian-4 theory. *J. Chem. Phys.* **126,** 084108 (2007).
37. Baboul, A. G., Curtiss, L. A., Redfern, P. C. & Raghavachari, K. Gaussian-3 theory using density functional geometries and zero-point energies. *J. Chem. Phys.* **110,** 7650 (1999).
38. Cossi, M., Rega, N., Scalmani, G. & Barone, V. Energies, structures, and electronic properties of molecules in solution with the C-PCM solvation model. *J. Comput. Chem.* **24,** 669–681 (2003).
39. Case, D. A. *et al.* Amber 11 (Univ. California, 2010).
40. Pellegrini, E. & Field, M. J. A generalized-Born solvation model for macromolecular hybrid-potential calculations. *J. Phys. Chem. A* **106,** 1316–1326 (2002).
41. Roberts, R. J. & Cheng, X. Base flipping. *Annu. Rev. Biochem.* **67,** 181–198 (1998).
42. Luo, Y. R. *Comprehensive Handbook of Chemical Bond Energies* Ch. 3 (CRC, 2007).



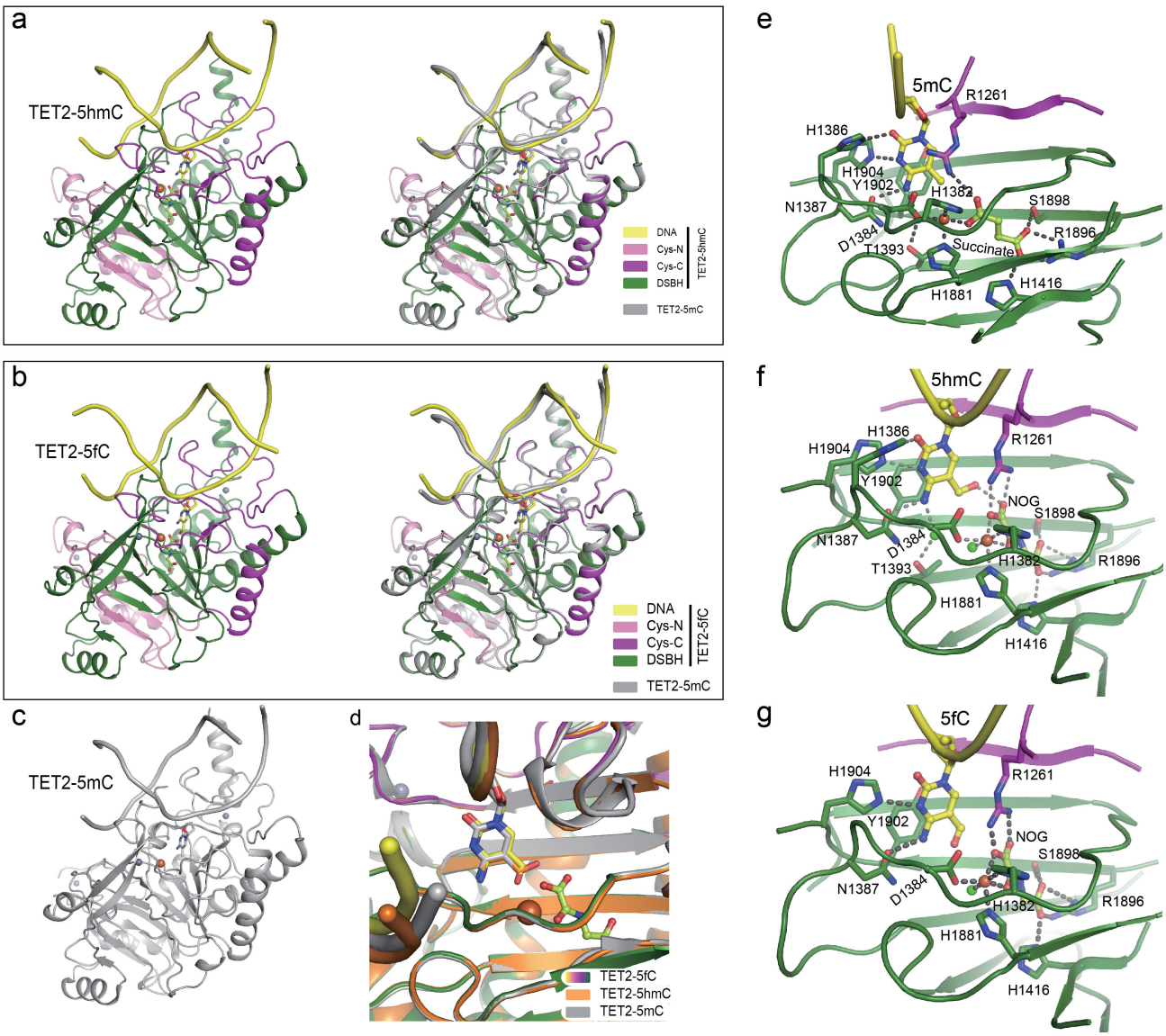
Extended Data Figure 1 | Steady-state kinetic analyses for TET-mediated oxidation on 5mC/5hmC/5fC-DNA substrates. a, Protein purification of human TET1 catalytic domain. Representative gel-filtration profile of human TET1 (residues 1418–2136) is shown. The peak position is about 13 ml, which corresponds to the monomer of TET1 with molecular mass of about 79 kilodaltons. The peak fractions were subjected to SDS–polyacrylamide gel electrophoresis and stained by Coomassie blue. The column used for gel filtration was Superdex 200 (GE Healthcare, 10/300 GL). **b**, Standard curves for 5mC, 5hmC, 5fC, 5caC, $[^2H]_3$5mC, $[^2H]_3$5hmC and guanine for quantification in LC–MS/MS. Good linearity was obtained for the range of guanine and various cytosine derivatives as indicated. The level of guanine (equal to total cytosine and its derivatives) was detected. Note that three standard curves were generated for 5mC/5hmC/5fC in low/mid/high concentrations, respectively. Two standard curves were generated for 5caC and $[^2H]_3$5mC in mid/low-concentration. **c–e**, Reaction progress curves of substrate and fraction products versus incubation time (2.5 min). Initial rates (nanomoles per second) for product generation were measured using various concentrations of dsDNA substrate and TET2. The reactions were conducted using 58-bp dsDNA (D5, one central 5mC/5hmC/5fCpG site) as substrate. Quantification was calculated from two independent assays; error bars, s.d. for duplicate experiments. **f–h**, To avoid generation of multiple products, we optimized protein concentration so that only one product was detected for all the reactions under our experimental conditions. The product generations are shown for the reactions with the highest substrate concentration for the longest time.



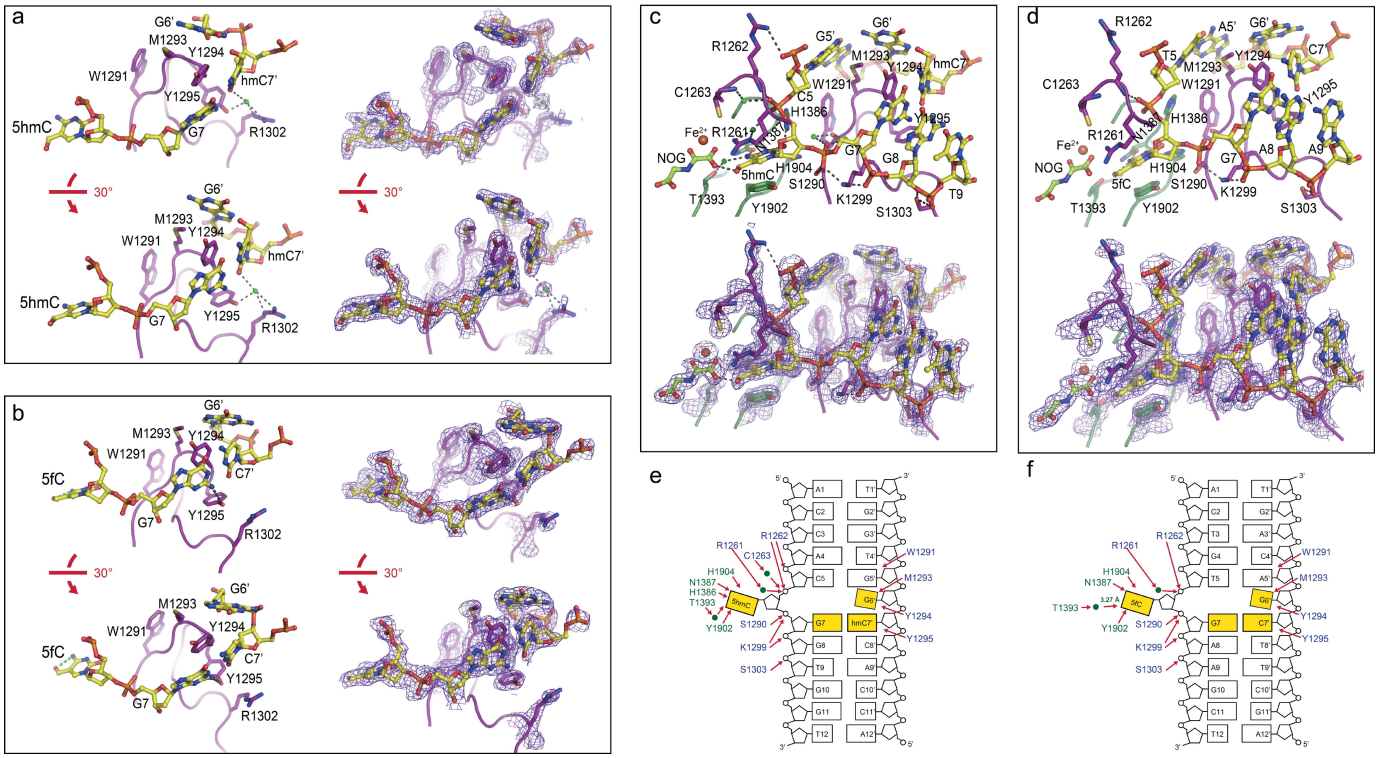
Extended Data Figure 2 | Similar substrate preference of TET2 towards DNA with different lengths and sequences. a–e, Enzymatic activity of TET2 was measured on 5mC/5hmC/5fC-DNA substrates (one central 5mC/5hmC/5fCpG site) of different lengths (**a**–**c**) or sequences (**b**, **d**, **e**). dsDNA substrates of 26 bp, 58 bp and 100 bp were used for reactions to test the effect of DNA length on substrate preference of TET2 (**a**–**c**). AT- and CG-rich dsDNA substrates of 58 bp were used for reactions to compare the effect of DNA sequence on substrate preference of TET2 (**b**, **d**, **e**). TET2 (1 μM) was used for all reactions. Quantification was calculated from three independent assays; error bars, s.d. for triplicate experiments. TET2 showed similar substrate preference (higher activity on 5mC-DNA than on 5hmC/5fC-DNA) for all DNA substrates tested. In addition, TET2 has no sequence preference apart from CpG dinucleotide. Notably, TET2 showed higher activity on shorter DNA substrate under our experimental conditions. It is not surprising because the enzyme should be able to find one CpG site easier on shorter DNA substrate (compared with longer DNA). **f**–**h**, Effects of substrate/product on enzymatic activities of TET2. **f**, Enzymatic activities of 1 μM TET2 for 26-bp 5mC-DNA in the presence of 26-bp 5C/5hmC-DNA. **g**, Enzymatic activities of 1 μM TET2 for 26-bp 5hmC-DNA in the presence of 26-bp 5C/5fC-DNA. **h**, Enzymatic activities of 1 μM TET2 for 26-bp 5fC-DNA in the presence of 26-bp 5C/5caC-DNA. The presence of CpG-DNA or different substrate/products has negligible impact on TET2 activity.



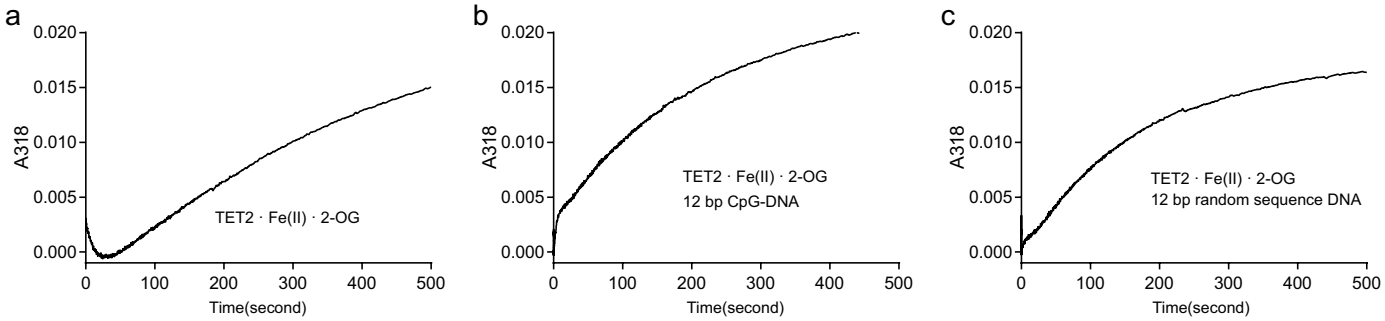
Extended Data Figure 3 | Fluorescence polarization measurements of DNA-binding affinities of TET2 in different conditions. a–c, Fluorescence polarization measurements of substrate DNA-binding affinities of TET2 (10 mM HEPES pH 7.0, 100 mM NaCl) in the presence of 100 μM Mn^{2+} and 1 mM 2-OG (**a**), 100 μM Fe^{2+} and 1 mM NOG (**b**) and 100 μM Fe^{2+} and 1 mM succinate (**c**). No significant difference was observed for the DNA-binding affinity of TET2 for different substrate/product under conditions mimicking oxidation or after oxidation.



Extended Data Figure 4 | Structural comparison of TET2–5mC-DNA, TET2–5hmC-DNA and TET2–5fC-DNA complexes. a–c, Structural comparison of the three complexes. The individual structures of the three complexes are shown on the left panel and the superimposed structures are shown on the right. The structures are shown in ribbon representations. TET2–5hmC-DNA and TET2–5fC-DNA are coloured as in Fig. 3b, and TET2–5mC-DNA is coloured in grey. The colour scheme is indicated. Stick representations show 5mC and 5hmC in two structures. **d**, Close-up view for the comparison of 5mC, 5hmC and 5fC in the three structures. Note that the cytosine portions of the three bases adopt almost identical conformations within the catalytic cavity. **e–g**, Close-up views of the catalytic DSBH core of the three structures, shown in ribbon representation with critical residues indicated in stick representations. The nitrogen and oxygen atoms are coloured in blue and red, respectively. Hydrogen bonds and Fe(II) coordination are indicated as dashed lines. Fe(II) and crystallographic water molecules are shown as red and green balls, respectively. The cytosine of 5hmC is specifically recognized by TET2 in a similar manner to that of 5mC in the TET2–5mC-DNA structure. An additional hydrogen bond between the hydroxymethyl group of 5hmC and NOG was observed in the TET2–5hmC-DNA structure. The additional hydrogen bond may not be strong enough to affect the binding affinity between TET2 and DNA because the interaction is mediated by extensive hydrogen bonds and hydrophobic interactions (Extended Data Fig. 5).



Extended Data Figure 5 | Recognition of CpG dinucleotide by TET2. **a**, Two different views of the interaction between the G7:hmC7′ base pair and TET2 for the specific recognition of CpG dinucleotide in TET2–5hmC-DNA. Water-mediated hydrogen bonds are formed between base hmC7′ of DNA and residues Y1295 and R1302 of TET2. The $2F_{observed} - F_{calculated}$ simulated annealing omit maps for residues involved in CpG recognition outside catalytic cavity are shown. The maps were calculated at 1.80 Å and contoured at 1.0σ. **b**, Two different views of the interaction between the G7:C7′ base pair and TET2 for the specific recognition of CpG dinucleotide in TET2–5fC-DNA. The $2F_{observed} - F_{calculated}$ simulated annealing omit maps were calculated at 1.97 Å and contoured at 1.0σ. **c**, **d**, Close-up views of the interactions between TET2 and 5hmC-DNA (**c**) and TET2 and 5fC-DNA (**d**). Critical bases of DNA and residues of TET2 for the interactions are shown in stick representations. Hydrogen bonds are indicated as dashed lines. Water molecule is shown as a green ball. The nitrogen, oxygen and phosphorus atoms are coloured in blue, red and orange, respectively. The $2F_{observed} - F_{calculated}$ simulated annealing omit maps for residues involved in DNA interaction are shown with these residues omitted from the calculation. Most residues are well covered by the electron density, indicating that these residues were built correctly in the structural model. **e**, **f**, Representation of intermolecular contacts between TET2 and 5hmC-DNA (**e**) and 5fC-DNA (**f**).



Extended Data Figure 6 | Kinetic traces from the reactions of TET2·Fe(II)·2-OG. Traces from the reactions in the absence (**a**) and presence (**b**) of unmethylated CpG-DNA or random DNA (**c**). The formation of catalytic intermediate for each reaction was monitored by stopped-flow absorption at 318 nm. No decay of catalytic intermediate was observed for any of the measurements. These analyses serve as a negative control for the assays shown in Fig. 4c.

Extended Data Table 1 | DNA-binding affinities of TET2

a.

	Kd(μM)			
FAM-18-bp dsDNA		100μM Fe^{2+} 1mM NOG	100μM Mn^{2+} 1mM 2-OG	100μM Fe^{2+} 1mM Succinate
5C	0.54±0.06	0.81±0.09	0.80±0.08	
5mC	0.99±0.15	1.34±0.18	0.76±0.06	
5hmC	0.41±0.07	0.97±0.13	0.60±0.04	0.48±0.02
5fC	0.56±0.10	0.62±0.09	0.53±0.05	0.54±0.02
5caC				0.25±0.01

b.

	5mC	5hmC	5fC
K_{a1}(1/Ms)	8.2E+4 ± 2.8E+3	9.9E+4 ± 2.1E+3	1.6E+5 ± 4.5E+3
K_{d1}(1/s)	8.0E-2 ± 2.9E-3	1.8E-1 ± 2.4E-3	2.8E-1 ± 5.0E-3
K_{a2}(1/s)	6.4E-3 ± 2.4E-4	2.8E-3 ± 5.4E-5	3.5E-3 ± 6.7E-5
K_{d2}(1/s)	1.5E-2 ± 3.6E-4	4.2E-3 ± 1.5E-4	3.7E-3 ± 1.3E-4
KD(M)	6.8E-7	1.1E-6	9.0E-7
Rmax(RU)	54.95	58.44	40.87
Chi^2(RU^2)	0.493	0.258	0.198

a, Fluorescence polarization measurements of DNA-binding affinities of TET2. FAM-labelled 18-bp DNA (5 nM) was incubated with increasing amounts of TET2 in a buffer containing 10 mM HEPES pH 7.0 and 100 mM NaCl. Mn^{2+} (100 μM) and 2-OG (1 mM), or 100 μM Fe^{2+} and 1 mM NOG, were added to mimic the catalytic condition. Fe^{2+} (100 μM) and succinate (1 mM) were added to mimic the product release process. The binding affinities were measured using fluorescence polarization. Error bars, s.d. for triplicate experiments from three independent experiments. **b**, SPR measurements of DNA-binding affinities of TET2. The sensorgrams were analysed using BIA evaluation software (Biacore). The response curves of various protein concentrations were fitted according to the two-state binding model described by the following equation.

$$A + B \underset{K_{d1}}{\overset{K_{a1}}{\rightleftarrows}} AB \underset{K_{d2}}{\overset{K_{a2}}{\rightleftarrows}} A * B$$

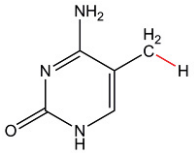
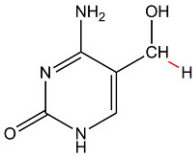
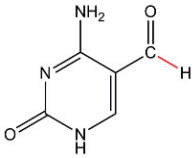
TET2 binds to 5mC, 5hmC or 5fC through a base flapping mechanism[9,41]. The process by which TET2 recognizes modified DNA is considered the first step, with the binding affinity calculated by the equation $K_{A1} = K_{a1}/K_{d1}$. The second step is base flipping, and the binding affinity is calculated by the equation $K_{A2} = K_{a2}/K_{d2}$. The overall equilibrium binding constant is calculated by the equation $K_A = K_{A1}(1 + K_{A2})$ and $K_D = 1/K_A$. The values of χ^2 for all three measurements are less than 1% R_{max}.

Extended Data Table 2 | Data collection and refinement statistics

	TET2-5hmC-DNA	TET2-5fC-DNA
Data collection		
Space group	C 2 2 2_1	C 2 2 2_1
Cell dimensions		
a, *b*, *c* (Å)	48.3, 87.5, 260.9	48.2, 88.0, 268.0
α, β, γ (°)	90.0, 90.0, 90.0	90.0, 90.0, 90.0
Resolution (Å)	50 - 1.80 (1.86 - 1.80) *	50 - 1.97 (2.04 - 1.97) *
R_{sym} or R_{merge}	0.094 (0.828)	0.055 (0.624)
$I/\sigma I$	19.2 (1.9)	28.0 (2.2)
Completeness (%)	98.0 (88.6)	98.9 (93.2)
Redundancy	9.6 (4.8)	6.0 (4.7)
Refinement		
Resolution (Å)	1.80	1.97
No. reflections	50686	57919
R_{work}/ R_{free}	0.177/0.212	0.205/0.248
No. atoms		
Protein	3259	3277
DNA	472	488
OGA	10	10
Ligand/ion	17	4
Water	309	262
B-factors		
Protein	43.4	31.7
DNA	63.1	47.5
OGA	33.0	23.5
Ligand/ion	51.6	27.5
Water	46.6	32.4
R.m.s deviations		
Bond lengths (Å)	0.006	0.009
Bond angles (°)	1.15	1.16

*Highest resolution shell is shown in parenthesis.

Extended Data Table 3 | Calculated C–H BDE for the 5-substitution groups of 5mC, 5hmC and 5fC

Base	C-H BDE (kcal/mol) at 298.15K			
	CBS-QB3	G4	G3(MP2B3)	CBS-QB3 (CPCM)
	89.74	88.46	91.01	90.39
	87.51	86.16	88.76	86.20
	91.98	89.02	91.89	92.89

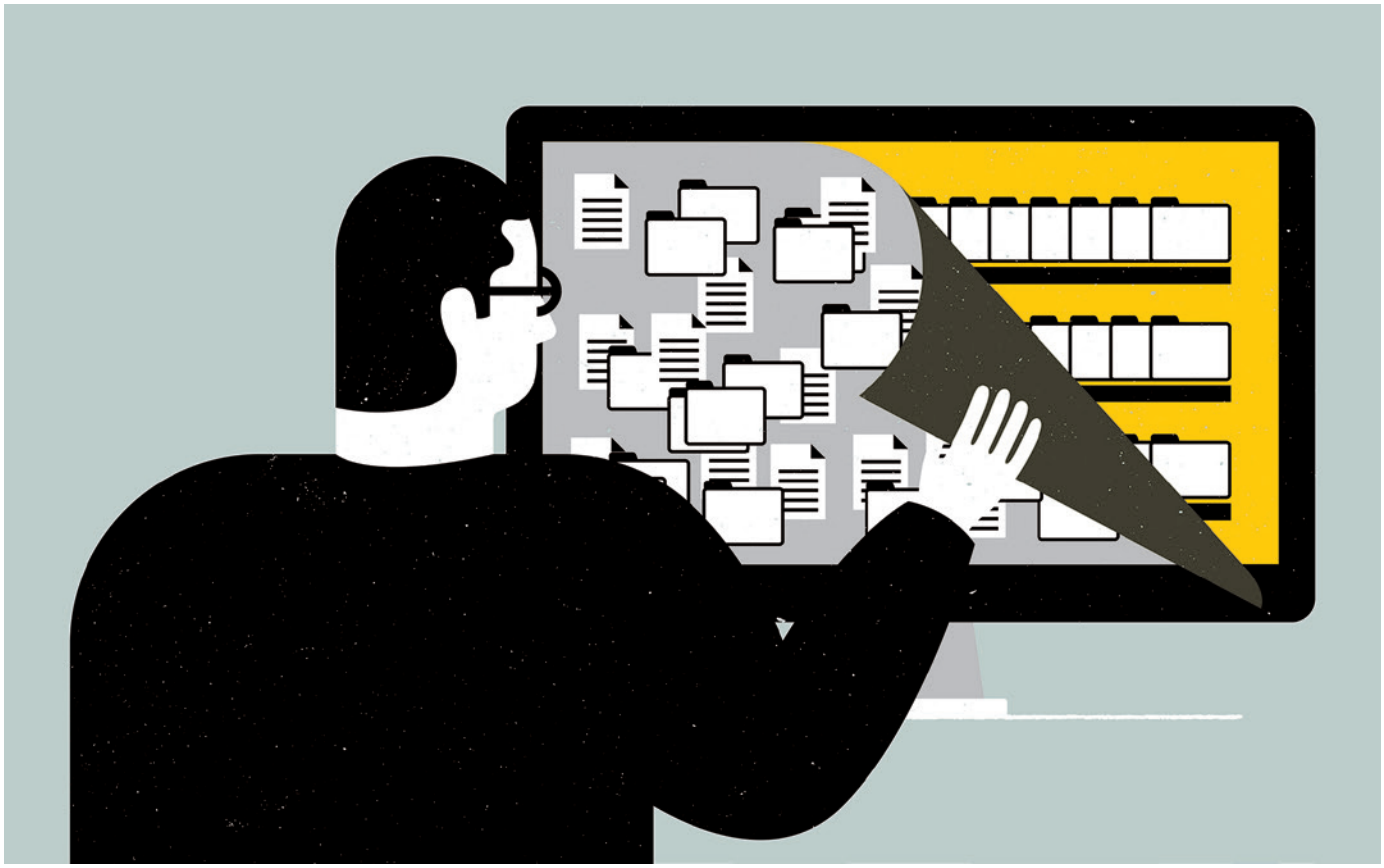
A bond cleavage reaction (R–H→R• + H•) was used to calculate each C–H BDE as the difference in enthalpies (ΔH) for the parent base (R–H) and the corresponding radical (R•) plus hydrogen atom (H•). Energies were calculated by using several high-precision computational methods implemented in Gaussian 09, including CBS-QB3, G4, G3(MP2B3) and CBS-QB3, with a C-PCM implicit water model. The experimentally estimated gas phase BDE for the corresponding C–H bonds in Ph-CH_3, Ph-CH_2OH and Ph-COH were reported to be 89.7 ± 1.2 kcal mol^{-1}, 79.0 ± 2.0 kcal mol^{-1} and 88.7 ± 2.6 kcal mol^{-1}, respectively[42].

TOOLBOX

EIGHT WAYS TO CLEAN A DIGITAL LIBRARY

Scientists have a surfeit of options to choose from in the competitive market of reference-management software.

ILLUSTRATION BY THE PROJECT TWINS



BY JEFFREY M. PERKEL

Adam Rocker didn't expect the software that managed his digital reference library to flag up better ways he could be doing his research. But his electronic filing system of choice, ReadCube, periodically scans his library and suggests related papers, rather as some music-file-management programs highlight recommended tunes. And that feature, he says, has brought up some unexpected gems.

As a graduate student, Rocker, who is now studying medicine at the University of Ottawa, was researching bacterial infections in zebrafish. ReadCube highlighted a paper that described a way to entrap the fish using microfluidics — a field whose literature he would not normally read — that was much easier than his own method. Being alerted to the research was "really rewarding", Rocker says, although he was ultimately too invested in his own project to adopt the alternative approach.

As Rocker discovered, today's reference-management tools go above and beyond simple electronic filing. Rather like a Swiss-army knife, each tool now appeals to customers by offering an ever-evolving set of extra features.

This article focuses on eight tools — colwiz, EndNote, F1000Workspace, Mendeley, Papers, ReadCube, RefME and Zotero — all competing in the reference-management market (see 'Reference-management software'). Some excel at streamlining the process of browsing and building literature libraries, whereas others focus on creating bibliographies, aiding collaboration through the use of shared workspaces or recommending papers. (One, ReadCube, is owned by Digital Science, a firm operated by the Holtzbrinck Publishing Group, which also has a share in *Nature*'s publisher.)

Each tool exists to help researchers to tame the digital flotsam and jetsam of scattered, downloaded PDFs. Most scientists can relate to that problem: as they grab PDFs from journal websites — where they are often assigned impenetrable alphanumeric codes as filenames — and dump them into any convenient folder, chaos can quickly take hold, with multiple ▶

copies of files spread across hard disks.

"In science, or at least in my experience, we tend to end up with a folder in the desktop with 3,000 really weirdly named PDF files, which we can never find when we need them," says Raúl Delgado-Morales, a neuroscientist at the Bellvitge Biomedical Research Institute in Barcelona, Spain.

Reference-management tools address that confusion by indexing a hard disk. Typically, the process of dragging and dropping a PDF into an application window triggers the software to try to identify it using the DOI or title, and to retrieve relevant metadata (such as title, keyword and author names) from online servers.

Researchers can also assign software to monitor specific folders into which they drop their files. They can then find PDFs through a simple search for author name, keyword or, in some cases, their own notes. Delgado-Morales solved his problem, for example, by organizing his literature library with Papers, a user-friendly application that automatically renames files according to any scheme he chooses. Other tools offer similar functions, except for RefME — a website and mobile app — which stores only lists of references and not the PDFs themselves.

REFERENCE-MANAGEMENT SOFTWARE
Eight of the most popular tools.

Product	URL	Platform	Free?
colwiz	colwiz.com	Desktop/web/mobile	Yes
EndNote	endnote.com	Desktop/web/mobile	Yes, with some limited features
F1000Workspace	f1000.com/work/	Web	No
Mendeley	mendeley.com	Desktop/web/mobile	Yes, with some limited features
Papers	papersapp.com	Desktop/web/mobile	No
ReadCube	readcube.com	Desktop/web	Yes, with some limited features
RefME	refme.com	Web/mobile (only stores references)	Yes
Zotero	zotero.org	Desktop/web/mobile	Yes, with some limited features

See the online version of this article at go.nature.com/xbp9ot for a fuller comparison.

CORE FUNCTIONS

Most of the tools help researchers to import literature from a variety of online sources. Many offer in-app searching of external databases such as PubMed and Google Scholar, as well as web-browser plugins that grab reference data (and sometimes, associated PDFs) from journal websites and other pages.

"We tend to end up with a folder with 3,000 really weirdly named PDF files."

Zotero — a free, open-source software project — was founded ten years ago specifically to tackle the problem of extracting information from a web browser, says project director Sean Takats of George Mason University in Fairfax, Virginia. "That's the key feature of Zotero, and remains one of its strongest compared to other reference managers," he says. RefME offers the unusual option of adding references by scanning a barcode with a smartphone camera.

One of the best-known features of reference-management software is the ability to insert in-text references in a research paper and to create bibliographies in any format. EndNote, a widely used commercial package, has offered this feature for decades, but now faces competition from many modern tools.

Many tools interface with common word-processing software (usually Microsoft Word, but sometimes OpenOffice and related freeware suites as well) so that a user typing up a research article need only select the papers that they want to mention and click a button to have codes inserted into the document to mark the in-text reference. Later, the user can create a bibliography and in-text citations according to several thousand journal styles, picking his or her choice from a pull-down list.

Most tools include built-in PDF readers for reading and annotating articles — typically allowing users to search through comments and notes — as well as cloud-based capabilities for syncing those comments (and the PDFs themselves) between, for example, an iPad and a desktop computer. But ReadCube and colwiz try to offer richer PDF reading experiences. In ReadCube, for instance, in-line citations and author names in PDFs are rendered as active hyperlinks to provide direct access to cited articles and publication lists. The same functionality is available when viewing and annotating PDFs on the websites of partnering publishers (including, for ReadCube, *Nature* and Wiley; and, for colwiz, Taylor & Francis).

Many of these tools can identify articles related to specific items in a library, or recommend articles on the basis of the library's content overall. F1000Workspace — like ReadCube — uses an algorithm to do this. It also taps into recommendations made by a community of 10,000 or so specialists. However, many other stand-alone software products also recommend papers (see *Nature* **513,** 129–130; 2014).

SET TO SHARE

Many tools now allow researchers to set up group libraries or share key papers with distant collaborators, although this process is carefully managed to prevent violation of publishers' copyright. Those in public groups using Mendeley, for instance, can share only information about a paper — the equivalent of a library-catalogue entry. Only users in private groups can share and modify PDFs (and groups must upgrade to a paid account to add more than three individuals).

Brenton Wiernik, an organizational-psychology PhD candidate at the University of Minnesota in Minneapolis, uses a shared library in Zotero for collaborative projects involving systematic reviews and meta-analyses of the literature in his field. Such efforts might involve 15–20 people, he says: some downloading articles into a shared library; others reading them; still more adding annotations and tags and logging key data.

According to Wiernik, the process is akin to using a shared Dropbox folder, with the added benefit that Zotero tracks and maintains metadata, notes and annotations. For instance, researchers can use a dedicated tag to indicate that they are processing an article, thereby signalling to collaborators that they should work on a different article to avoid duplicated effort.

F1000Workspace and colwiz both extend sharing to include features for preparing manuscripts and managing projects. With F1000Workspace, researchers can use a plugin to upload Microsoft Word manuscripts to a secure location, thereby enabling team members to comment on the shared copy — although the text cannot be edited in the browser, says João Peres, the company's product-development manager. Peres plans to implement a 'one-click' article-submission feature that sends papers directly from F1000Workspace to journal editors, starting with the journal *F1000Research*. And colwiz also permits users to share documents to an online drive for team members to view and comment on.

Given the highly overlapping feature sets of these tools, a user's choice often comes down to particular individual priorities. Richard Karnesky, a materials scientist at the Sandia National Laboratories in Livermore, California, supports Zotero for its open-source ethos, for example.

Perhaps the best reason for using a reference manager is the technology's ability to provide a form of searchable memory. Imagine, says Boyd Steere, a senior research scientist at pharmaceutical firm Eli Lilly in Indianapolis, Indiana, a desk piled high with printed papers: Post-it notes hanging out, writing in the margins, doodles, notations, arrows and more. Today's PDF-filled, digital folders are in many ways no easier to navigate. With a digital reference manager, however, buried knowledge is just a keyword search away. ■

Jeffrey M. Perkel *is a writer based in Pocatello, Idaho.*

SCIENCE COMMUNICATION The art, the practice, the opportunities **go.nature.com/qm5r6i**

DATA SHARING Why it does not happen **go.nature.com/bbtrzx**

NATUREJOBS For the latest career listings and advice **www.naturejobs.com**

ADAPTED FROM ER TEN HONG/GETTY

SUPERVISION

Clear direction

Managing laboratory members as well as a research strategy can be difficult for early-career principal investigators, but help is at hand.

BY BOER DENG

Vivek Kumar admits that he has not always been the best manager. Routinely, the neuroscientist would fail to provide important details about his expectations to junior colleagues, then lose his temper when they did not meet those expectations. In the laboratory where he conducted his postdoctoral research, for example, Kumar tasked the technician with cloning cells but did not give her a deadline. She had not completed the work when he demanded the clones, and she later told him that her blood pressure would rise whenever she heard him approaching.

The comment might have been difficult to hear, but it helped Kumar to realize that he needed to improve his management skills. When he set up his own lab in January 2015 at the Jackson Laboratory in Bar Harbor, Maine, he was determined to receive training in how to be a good leader, mentor and manager. A few months later, Kumar attended a workshop on leadership at the Cold Spring Harbor Laboratory in New York. There, he learned about the communication and negotiation skills that would help him in his role as principal investigator (PI). But almost one year on, that role can still feel uncomfortable. Managing people remains one of his biggest challenges, Kumar acknowledges — especially when it comes to having difficult conversations with colleagues about expectations. However, the course did teach him new skills and tactics. "I came away from the workshop with a clear sense that it's part of my responsibility to make the whole lab a success."

Many junior researchers say that they feel poorly prepared for managerial roles. "Knowing how to do good science, that's the price of admission for being a researcher," says Jeff Gustafson, an organic chemist who has led a lab ▶

for three years at San Diego State University in California. "But when I started my own lab, there were other things that I just had no idea how to do." Juggling the challenges of teaching and administrative duties while guiding the members of his lab was a mixture for which he had not been prepared.

Graduate students, junior researchers and their institutions have been awakened to the fact that, early in their careers, they need to develop the interpersonal skills that lab leaders require. "Over the past ten years, the interest in learning management as scientists has gone from a trickle to a small stream," says Carl Cohen, an executive coach for scientists who, in 2011, helped to start the leadership programme that Kumar attended at the Cold Spring Harbor Laboratory. In fact, a number of institutions have launched workshops and seminars to teach management to postdoctoral researchers and junior faculty members (see 'Learn to lead').

One reason for the increase in management-training options for early-career researchers is that although universities are producing more researchers, many will not remain in academia. Former trainees often enter fields in which management skills comprise a significant component of their jobs. "Students and their PIs know that they may not have the same careers," says Cohen, who taught and led research in molecular haematology at Tufts University in Medford, Massachusetts, before holding executive positions at several biotechnology companies.

AVOID CONFLICT

Academic scientists have also realized the importance of good management for success. For example, it is easier to attract talented researchers to a lab that has no conflicts, points out Markus Seeliger, who leads a cancer and ageing research group at Stony Brook School of Medicine in New York. Junior faculty members can highlight this selling point to potential recruits, who might otherwise want to work for more established researchers.

"Over the past ten years, the interest in learning management as scientists has gone from a trickle to a small stream."

Kathy Barker, a microbiologist turned author and management consultant in Seattle, Washington, has noticed that an increasing number of scientists now mentor each other and address the cultural and interpersonal aspects of science. "In the first lab I worked in, no one talked to me for three days because I asked the wrong person how to use the autoclave," recalls Barker, who in 2001 published *At the Helm* (Cold Spring Harbor Laboratory Press), a management guidebook for inexperienced PIs. Her experience spurred her to write about the importance of management and crafting a comfortable culture in which to do science. These days, many institutions pay attention to making their labs more welcoming, she says.

The field of research, number of members and culture of each lab bring their own predicaments for new PIs. "Issues can be quite different depending on whether you are working in a narrow field versus a field with lots of collaborative projects," says Justin Cotney, a developmental biologist at the University of Connecticut Health Center in Farmington. In small labs, interpersonal relationships between PIs and lab members are often more important — and potentially thorny — than in larger labs. Because PIs are able to spend more time and work more closely with postdocs and students in a small group, issues such as a communication problem or something not working are harder to ignore.

PIs can help by setting expectations and developing lab protocols that make negative feelings less likely to crop up. A month or two after setting up his lab at Georgetown University Medical Center in Washington DC, neuroscientist Patrick Forcelli received complaints from his disgruntled lab manager, who was upset about mess left in the lab and incomplete paperwork. Forcelli has since assigned a specific responsibility for lab upkeep to each member of his group, and devotes the beginning of the lab's weekly meetings to reviewing whether tasks have been completed. Making lab members accountable to each other has united everyone behind a shared standard — and has also made the lab a nicer place to work.

But sometimes the problems are not so easy to fix. As in any other workplace, the personalities and moods of individuals affect the overall lab environment. PIs must be attuned to how each member behaves in and perceives the work environment. "Knowing the people you work with and figuring out what each member of the lab will respond to helps you to know when a conflict might arise or escalate," says Cotney. He learned the lesson firsthand while he was a postdoc. When a colleague who had been struggling with personal issues snapped at a new junior researcher, Cotney stepped in to defuse the tension. He reminded his colleague not to direct unreasonable anger at another lab member. "It was good to be proactive, and is something I do as a PI." Forcelli says that in small labs, it is especially important for PIs to play an active part in handling conflicts. "I've seen cases where the PI will just be hands-off, which makes the environment miserable for several people in the lab for an indefinite period of time," he adds.

Kumar thinks that training can help researchers to appreciate the importance of good management. He says that the workshop he attended helped him to better understand his role and responsibilities. For PIs like Kumar, it can be a relief to know that they can learn discrete skills for resolving management challenges. Perhaps the most important lesson is learning to view difficulties as normal and tractable. "One thing I take away is that it's OK that something falls through — that you don't have to be perfect all the time. You realize that everybody is facing these things," says Cotney. "It's nice to know you're not alone." ■

Boer Deng, *a former* Nature *intern, is the Washington DC correspondent for* The Times.

LEARN TO LEAD

Management resources abound

Management science has existed for more than a century. In 1911, engineer Frederick Taylor outlined the principles of 'scientific management', which aims to improve productivity in the workplace through collaboration. Management resources for early-career researchers are increasing. Here are a few.

- The Leadership in Bioscience workshop at the Cold Spring Harbor Laboratory in New York runs for 3.5 days every February or March. Aimed at postdoctoral researchers who are about to take leadership of a lab, as well as early-career principal investigators, the workshop accepts around 25 students, from a pool of about 40 applicants.
- The European Molecular Biology Organization (EMBO) in Heidelberg, Germany, holds a comprehensive series of workshops for early-career scientists. When they began in 2005, the workshops were offered only five or six times a year. Now, they take place 20 times a year, with each workshop of 16–20 participants filling quickly. There is a waiting list for EMBO's lab-management courses for principal investigators and postdoctoral researchers.
- The UK-based Vitae online resource offers career-development advice for researchers. Registered members around the world can access tools to learn about conflict management and coaching for researchers, as well as other areas of professional growth.
- The Jackson Laboratory in Bar Harbor, Maine, offers a course called The Whole Scientist, which helps graduate and postdoctoral researchers to make the leap from acolyte to doyen. Georgetown University in Washington DC holds a similar course for early-career researchers.
- And this year, the Van Andel Research Institute in Grand Rapids, Michigan, began a series of workshops in leadership and management skills for scientists that it plans to continue yearly. **B.D.**

nature OUTLOOK

5 November 2015
Supplement to Nature Publishing Group journals

BIG DATA IN BIOMEDICINE

Produced with support from:

PHOENIX

BPRC
Beijing Proteome Research Center

Harnessing the information explosion

ONE SLOW STEP FOR MAN

Survival instinct.

BY S. R. ALGERNON

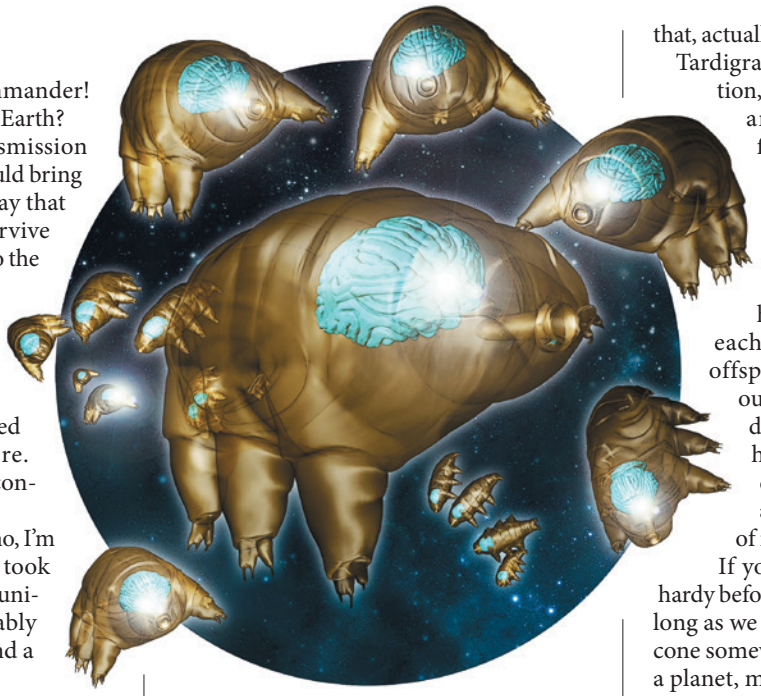


ILLUSTRATION BY JACEY

Greetings, Mission Commander! How are things back on Earth?

I wish our first transmission from the α-Centauri system could bring better news, but I'm sorry to say that Captain Thurgood did not survive the trip. Something happened to the CO_2 filters, I'm afraid. The rest of the crew died as well, some sooner than others. That's what the logs say, anyway. As for the details … well, you might as well ask me what caused the fall of the Roman Empire. Questions on that scale don't concern us much anymore.

In case you were wondering, no, I'm not part of the crew. In fact, it took years to drift over to the communications console and considerably longer to figure out how to send a transmission.

You probably don't know me. We might have passed in the hall back at the lab, before your project squeezed ours out entirely. Once, my name used to be on your office door.

We designed nanocomputer components. Ours were the best in the business, an order of magnitude smaller than anybody else's. That turned out to be the problem. Computers were small enough, now, the committee said. Once you can squeeze a petabyte onto a grain of sand, they said, you can do just about anything. Humanity has no need for computers that small.

I knew I could prove them wrong, if I gave it a little thought. That was when I read your press releases and noticed the biological samples that were part of your interstellar payload. I searched the Internet for 'tardigrade', and I saw my chance.

Dr Ehrlinger had found work on your team monitoring the life signs of the biological samples … for a substantial pay cut, I should add. I met her for drinks in a diner across from the launch facility, and the plan fell into place.

Tardigrade means 'slow-stepper'. Some people prefer 'water bear', maybe because they don't want anybody to think they're slow, but they're in no hurry. They've been ambling about for half-a-billion years or so. Humans don't faze them one bit. Tardigrades can survive just about anywhere, even in deep space.

Tardigrades are a millimetre or so in length, so I don't have quite the same stature I once did. I've grown, though, in my own way.

Maybe I lied before. Maybe I'm not the same person whose office you poached, not exactly. I'm a neurocognitive simulation that fits inside a 0.1-mm brain case. I might not have all the wetware of the original, but I've got it where it counts. The human brain has a lot of redundancy. It's amazing what you can do when you really get serious about compression. We even had enough room for a little 3D printer on one end, for enhancements and self-replication.

My entire team is here, including Dr Ehrlinger, at least in their computerized forms. Our human versions are still on Earth, on some quiet little island, somewhere where they don't extradite. It wasn't too hard to smuggle ourselves aboard with our tardigrade hosts. Once we trained them to move the way we wanted them to, we had the run of the ship.

I know what you're thinking. We called you to gloat about tanking your mission. None of us are pilots. The ship is going to burn up on re-entry anyway, so who cares if it's infested by a bunch of vindictive machines and wayward 'bugs'?

We're fine with that, actually. We can take it.

Tardigrades can handle vacuum, radiation, desiccation, heat and just about anything else. Besides, we've figured out how to manipulate the somatic and germline tissues of our hosts. We've been pushing them to reproduce and spurring their evolution. It's thrilling, actually, to herd the sperm and egg towards each other, creating just the right offspring, and then to bury one of our newly replicated brains in the developing embryo. Dr Ehrlinger has a knack for genetics, and we don't need our human bodies any more to appreciate the joys of reproduction.

If you think the tardigrades were hardy before, you ain't seen nothing yet. As long as we can manage to point the nose-cone somewhere in the neighbourhood of a planet, most of us will get through with barely a hiccup.

Isn't that great news? You can tell everyone at Mission Control that you've succeeded beyond your wildest expectations. You can take all the credit if you want. All that matters is that we have a home now, and a sense of purpose, and a plan for the future. We never could have done it without you.

In fact, most of us don't even hold a grudge any more. I have to admit that my programming was crude at the outset, and revenge was a fixation of mine. Our machine-learning algorithms and the chunks of code we've swapped with each other over countless generations have broadened our horizons.

I like to think that we've grown as far beyond you in the past few decades as you have in the past hundred million years of evolution. Maybe I'm underestimating you, though. When you get here, we'll find out who's smarter than whom.

Take your time. Slow and steady. That's the tardigrade way.

We're a patient lot. When you do arrive, you'll find us rather laid-back and democratic. One sentient organism, one vote, and all that.

Just don't be surprised if by then we outnumber you by a trillion or so to one. ■

NATURE.COM
Follow Futures:
@NatureFutures
go.nature.com/mtoodm

S. R. Algernon *studied fiction writing and biology, among other things, at the University of North Carolina at Chapel Hill. He currently lives in Singapore.*

nature OUTLOOK

BIG DATA IN BIOMEDICINE

5 November 2015 / Vol 527 / Issue No 7576

Cover art: Tatiana Plakhova

It may now cost less to sequence the three billion DNA base pairs of a human genome than to do a brain scan. But how does all that genomic data translate into treatment?

Life scientists are bringing together astonishing volumes of information from genomic sequencing, lab studies and patient records. And the resulting era of 'precision medicine' is already delivering treatments tailored to individual needs.

These 'big data' efforts face huge challenges, from creating analytic tools and solving scientific puzzles to accessing millions of gigabytes of data and overcoming barriers to accessing patients' health records (see pages S2 and S19).

Dozens of international projects are producing huge amounts of biomedical information, not just on the genome but on many other '-omes' (S8). Giant strides are being made in mapping the human proteome and building a 'parts list' of the body (S6). Meanwhile, smartphones and other wearable devices are generating continuous flows of health data from large numbers of people (S12). This vast array of data will allow a more detailed understanding of disease traits in analyses known as deep phenotyping (S14). Research organizations are assembling cloud-based 'information commons' to standardize, store and share the data (S16).

Drug companies are facing complex choices (S18). Many are opting to treat cancer, a main thrust in national programmes such as the UK 100,000 Genomes Project (S5). And some of these therapies are already changing clinical practice (S10).

We are pleased to acknowledge that this Outlook was produced with support from the National Center for Protein Sciences–Beijing, Beijing Proteome Research Center, State Key Laboratory of Proteomics, China Human Proteome Organization, Beijing Institute of Radiation, and the Academy of Military Medical Sciences. As always, *Nature* retains sole responsibility for all editorial content.

Eric Bender
Contributing Editor

Nature Outlooks are sponsored supplements that aim to stimulate interest and debate around a subject of interest to the sponsor, while satisfying the editorial values of *Nature* and our readers' expectations. The boundaries of sponsor involvement are clearly delineated in the Nature Outlook Editorial guidelines available at go.nature.com/e4dwzw

CITING THE OUTLOOK
Cite as a supplement to *Nature*, for example, *Nature* **Vol. XXX**, No. XXXX Suppl., Sxx–Sxx (2015).

VISIT THE OUTLOOK ONLINE
The *Nature Outlook Big Data in Biomedicine* supplement can be found at http://www.nature.com/nature/outlook/big-data
It features all newly commissioned content as well as a selection of relevant previously published material.

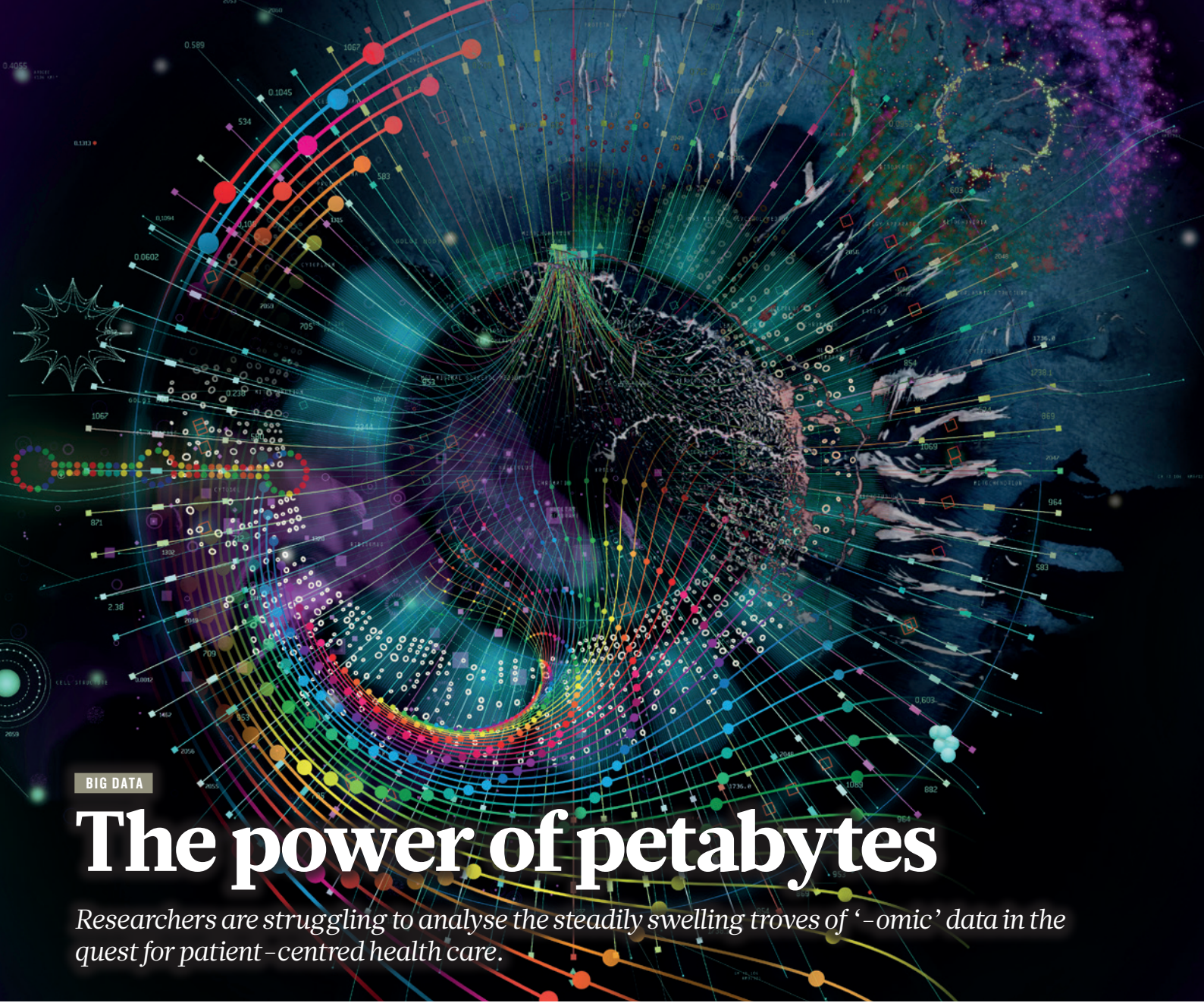
All featured articles will be freely available for 6 months.

SUBSCRIPTIONS AND CUSTOMER SERVICES
For UK/Europe: Nature Publishing Group, Subscriptions, Brunel Road, Basingstoke, Hants, RG21 6XS, UK. Tel: +44 (0) 1256 329242.
Subscriptions and customer services for Americas – including Canada, Latin America and the Caribbean: Nature Publishing Group, 75 Varick St, 9th floor, New York, NY 10013-1917, USA. Tel: +1 866 363 7860 (US/Canada) or +1 212 726 9223 (outside US/Canada). Japan/China/Korea: Nature Publishing Group — Asia-Pacific, Chiyoda Building 5-6th Floor, 2-37 Ichigaya Tamachi, Shinjuku-ku, Tokyo, 162-0843, Japan. Tel: +81 3 3267 8751.

CUSTOMER SERVICES
Feedback@nature.com

CONTENTS

RELATED RESEARCH



BIG DATA

The power of petabytes

Researchers are struggling to analyse the steadily swelling troves of '–omic' data in the quest for patient–centred health care.

BY MICHAEL EISENSTEIN

Fifteen years ago, it was a landmark achievement. Ten years ago, it was an intriguing but highly expensive research tool. Now, falling costs, soaring accuracy and a steadily expanding base of scientific knowledge have brought genome sequencing to the cusp of routine clinical care.

A growing number of institutions are conducting genome-wide 'dragnet' searches to identify the mutations responsible for rare diseases. "The rate at which we're finding causative variants in those cases is going up," says Russ Altman, a bioinformatician at Stanford School of Medicine in California. "At some centres, it's up to 50% of cases." Genomic variants can also reveal 'driver' mutations that might reveal a tumour's therapeutic vulnerabilities, or provide clues to whether a specific individual may or may not respond to a drug — the drug's 'pharmacogenetic' properties.

The US$1,000 genome, initially conceived as a price point at which sequencing could become a component of personalized medicine, has arrived. "Our capacity for data generation relative to price has increased in a way that is almost unprecedented in science — roughly six orders of magnitude in the past seven or eight years," says Paul Flicek, a specialist in computational genomics at the European Molecular Biology Laboratory's European Bioinformatics Institute in Cambridge, UK. The HiSeq X Ten system developed by Illumina of San Diego, California, can sequence more than 18,000 human genomes per year, for example.

The biomedical research community is diving in whole-heartedly, with population-scale programmes that are intended to explore the clinical power of the genome. In 2014 the United Kingdom launched the 100,000 Genomes Project, and both the United States (under the Precision Medicine Initiative) and China (in a programme to be run by BGI of Shenzhen) have unveiled plans to analyse genomic data from one million individuals.

Many other programmes are under way that, although more regional in focus, are still 'big data' operations. A partnership between Geisinger Health System, based in Danville, Pennsylvania, and biotech firm Regeneron Pharmaceuticals of Tarrytown, New York, for instance, aims to generate sequence data for more than 250,000 people. Meanwhile, a growing number of hospitals and service providers worldwide are sequencing the genomes of people with cancers or rare hereditary disorders (see 'DNA sequencing soars').

Some researchers worry that the flood of data could overwhelm the computational pipelines needed for analysis and generate unprecedented demand for storage — one article estimated that the output from genomics may soon dwarf data heavyweights such as YouTube. Many also worry that today's big data lacks the richness to provide clinical value. "I don't know if a million genomes is the right number, but clearly we need more than

TATIANA PLAKHOVA

we've got," says Marc Williams, director of the Geisinger Genomic Medicine Institute.

THE MEANING OF MUTATIONS

Clinical genomics today is largely focused on identifying single-nucleotide variants — individual 'typos' in the genomic code that can disrupt gene function. And rather than looking at the full genome, many centres focus instead on the exome — the subset of sequences containing protein-coding genes. This reduces the amount of data being analysed nearly 100-fold, but the average exome still contains more than 13,000 single-nucleotide variants. Roughly 2% of these are predicted to affect the composition of the resulting protein, and finding the culprit for a given disease is a daunting challenge.

For decades, biomedical researchers have dutifully deposited their discoveries of single-nucleotide variants in public resources such as the Human Gene Mutation Database, run by the Institute of Medical Genetics at Cardiff University, UK, or dbSNP, maintained by the US National Center for Biotechnology Information. However, the effects of these mutations were often determined from cell culture or animal models, or even theoretical predictions, providing insufficient guidance for clinical diagnostic tools. "In many cases, associations were made with relatively low levels of evidence," says Williams.

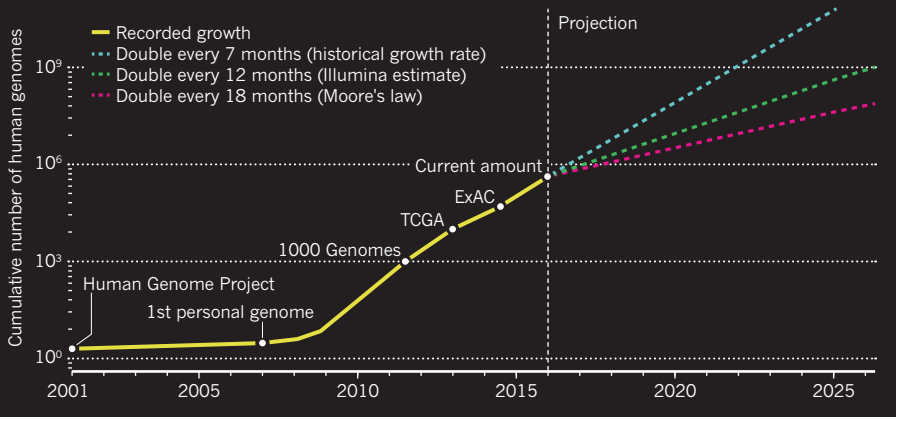
The situation is even more complicated for structural variants, such as duplicated or missing chunks of genome sequence, which are far more difficult to detect with existing sequencing technologies than single-nucleotide variants. At the whole-genome scale, each person has millions of variants. Many of these are in sequences that do not encode proteins but instead regulate gene activity, so they can still contribute to disease. However, the extent and function of these regulatory regions are poorly defined. Although capturing all this variability is desirable, it may not offer the best short-term returns for clinical sequencing. "You're shooting yourself in the foot if you're collecting data you don't know how to interpret," says Altman.

Efforts are now under way to rectify this problem. The Clinical Genome Resource, which was set up by the US National Human Genome Research Institute, is a database of disease-related variants, and contains information that could guide medical responses to these variants as well as the evidence supporting those associations. Genomics England, which runs the 100,000 Genomes Project, aims to bolster progress in this area by establishing 'clinical interpretation partnerships': doctors and researchers will collaborate to establish robust models of diseases that can potentially be mapped to specific genetic alterations.

"You're shooting yourself in the foot if you're collecting data you don't know how to interpret."

DNA SEQUENCING SOARS

Human genomes are being sequenced at an ever-increasing rate. The 1000 Genomes Project has aggregated hundreds of genomes; The Cancer Genome Atlas (TGCA) has gathered several thousand; and the Exome Aggregation Consortium (ExAC) has sequenced more than 60,000 exomes. Dotted lines show three possible future growth curves.



However, quantity is as important as quality. Mutations that offer a strong detrimental effect bring an evolutionary disadvantage, so they tend to be exceedingly rare and require large sample sizes to detect. Establishing statistically meaningful disease associations for variants with weak effects also needs large numbers of people.

In Iceland, deCODE Genetics has demonstrated the power of population-scale genomics, combining extensive genealogy and medical-history records with genome data from 150,000 people (including 15,000 whole-genome sequences). These findings have allowed deCODE to extrapolate the population-wide distribution of known genetic risk factors, including gene variants linked to breast cancer, diabetes and Alzheimer's disease.

They have also enabled studies in humans that normally require the creation of genetically modified animals. "We have established that there are about 10,000 Icelanders who have loss-of-function mutations in both copies of about 1,500 different genes," says Kári Stefánsson, the company's chief executive. "We're putting significant effort into figuring out what impact the knockout of these genes has on individuals."

This work was helped by the homogeneous nature of the Icelandic population, but other projects require a broadly representative spectrum of donors. Efforts such as the international 1000 Genomes Project have catalogued some of the world's genetic diversity, but most data are heavily skewed towards Caucasian populations, making them less useful for clinical discovery. "Because they come from the genetic mother ship, so to speak, people of African ancestry carry a lot more genetic variants than non-Africans," says Isaac Kohane, a bioinformatician at Harvard Medical School in Boston, Massachusetts. "Variants that seem unusual in Caucasians might be common in Africans, and may not actually cause disease."

Part of the problem stems from the reference genome — the yardstick sequence by which scientists identify apparent abnormalities, developed by the multinational Genome Reference Consortium. The first version was cobbled together from a few random donors of undefined ethnicity, but the latest iteration, known as GRCh38, incorporates more information about human genomic diversity.

INTO THE CLOUD

Harvesting genomes or even exomes at the population scale produces a vast amount of data, perhaps up to 40 petabytes (40 million gigabytes) each year. Nevertheless, raw storage is not the primary computational concern. "Genomicists are a tiny fraction of the people who need bigger hard drives," says Flicek. "I don't think storage is a significant problem."

A greater concern is the amount of variant data being analysed from each individual. "The computation scales linearly with respect to the number of people," says Marylyn Ritchie, a genomics researcher at Pennsylvania State University in State College. "But as you add more variables, it becomes exponential as you start to look at different combinations." This becomes particularly problematic if there are additional data related to clinical symptoms or gene expression. Processing data of this magnitude from thousands of people can paralyse tools for statistical analysis that might work adequately in a small laboratory study.

Scaling up requires improvisation, but there is no need to start from scratch. "Fields like meteorology, finance and astronomy have been integrating different types of data for a long time," says Ritchie. "I've been to meetings where I talk to people from Google and Facebook, and our 'big data' is nothing like their big data. We should talk to them, figure out how they've done it and adopt it into our field."

Unfortunately, many talented programmers with the skills to wrangle big data sets are lured away by Silicon Valley. Philip Bourne, associate director for data science at the US

National Institutes of Health (NIH), believes that this is partly due to a lack of recognition and advancement within a publication-driven system of scientific credit that leaves software creators and data managers out in the cold. "Some of these people truly want to be scholars, but they can't get the stature of faculty — that's just not right," says Bourne.

Processing power is another limiting factor. "This is not a desktop game — the real practitioners are proficient in massively parallel computation with hundreds if not thousands of CPUs, each with large memory," says Kohane. Many groups that analyse massive amounts of sequence data are moving to 'cloud'-based architectures, in which the data are deposited within a large pool of computational resources and can then be analysed with whatever processing power is required.

"There's been a gradual evolution towards this idea that you bring your algorithms to the data," says Tim Hubbard, head of bioinformatics at Genomics England. For Genomics England, this architecture is contained in a secure government facility, with strict control over external access. Other research groups are turning to commercial cloud systems, such as those provided by Amazon or Google.

PRIVACY PROTECTION

In principle, cloud-based hosting can encourage sharing and collaboration on data sets. But regulations on patient consent and privacy rights surrounding highly sensitive clinical information pose tricky ethical and legal issues.

In the European Union, collaboration is impeded by member states having different rules on data handling. Sharing with non-EU nations relies on cumbersome mechanisms to establish adequacy of data protection, or restrictive bilateral agreements with individual organizations. To help solve this problem, a multinational coalition, the Global Alliance for Genomics and Health, developed the Framework for Responsible Sharing of Genomic and Health-Related Data. The Framework includes guidelines on privacy and consent, as well as on accountability and legal consequences for those who break the rules.

"In data-transfer agreements, you could save yourself pages and pages of rules if the institution, researcher and funder agree to follow the Framework," says Bartha Knoppers, a bioethicist at McGill University in Montreal, Canada, who chairs the Alliance's regulatory and ethics working group. The Framework also calls for 'safe havens' that allow the research community to analyse centralized banks of genomic data that have been identity-masked but not fully 'de-identified', so they remain useful. "We want to link it to clinical data and to medical records, because we're never going to get to precision medicine otherwise, so we're going to have to use coded data," explains Knoppers.

Integrating genomics into electronic health records is becoming increasingly important for many European nations. "Our objective is to put this into the standard National Health Service," says Hubbard. The UK 100,000 Genomes Project may be the furthest along at the moment, but other countries are following. Belgium recently announced an initiative to explore medical genomics, for example.

LEN RUBENSTEIN/THE BROAD INSTITUTE

Rapid advances in technology are transforming genomics research.

All these nations benefit from having centralized, government-run health-care systems. In the United States, the situation is more fragmented, with different providers relying on distinct health-record systems, supplied by different vendors, that are generally not designed to handle complex genomic data. The NIH launched the Electronic Medical Records and Genomics (eMERGE) Network in 2007 to define best practices.

FROM DATA TO DIAGNOSIS

The immediate goal of genomically enriched health records is to explain the implications of gene variants to physicians, and one of its earliest implementations is pharmacogenetics. The Clinical Pharmacogenetics Implementation Consortium has translated known drug–gene interactions reported in PharmGKB (a database run by Altman and his colleagues) for clinical use. For example, people with certain variants may respond poorly to particular anticoagulants, leading to increased risk of heart attack. "The issue there is, how do you take a practitioner who has 12 minutes per patient and about 45 seconds of time allocated for prescribing drugs, and influence their practice in a meaningful way?" says Altman.

As long as deciding how to adapt care to genetic findings remains a job for humans, this process will remain time- and labour-intensive. Nevertheless, combining genotype and phenotype information is proving fruitful from a research perspective. Most clinically relevant gene variants were identified through genome-wide association studies, in which large populations of people with a given disease were examined to identify closely associated genetic signatures. Researchers can now work backwards from health records to determine what clinical manifestations are prevalent among individuals with a given genetic variant.

And the genome is only part of the story — other '-omes' may also be useful barometers of health. In July, Jun Wang stepped down as chief executive of BGI to start up an organization to analyse BGI's planned million-genome cohort alongside equivalent data sets from the proteome, transcriptome and metabolome. "I will be initiating a new institution to focus on using artificial intelligence to explore this kind of big data," he says.

IT TAKES PATIENTS

As researchers strive to integrate data from health records and clinical trials with genomic and other physiological data, patients are starting to contribute. "When we're focused on things like behaviour, nutrition, exercise, smoking and alcohol, you can't get better data than what patients report," says Ritchie.

Wearable devices, such as smartphones and FitBits, are collecting data on exercise and heart rate, and the volume of such data is soaring (see 'page S12) as it can be gathered with minimal effort on the wearer's part.

Each patient may become a big-data producer. "The data we generate at home or in the wild will vastly exceed what we accumulate in clinical care," says Kohane. "We're trying to create these big collages of different data modalities — from the genomic to the environmental to the clinical — and link them back to the patient." As these developments materialize, they could create computational crunches that will make today's 'big data' struggles seem like pocket-calculator problems. And as scientists find ways to crunch the data, patients will be the ultimate winners. ■

Michael Eisenstein *is a freelance science writer based in Philadelphia, Pennsylvania.*

KATE WITKOWSKA/GENOMICS ENGLAND

Q&A Mark Caulfield
National genomics

Mark Caulfield is chief scientist at Genomics England, which was set up in 2013 to deliver the UK 100,000 Genomes Project, initially focusing on cancers, rare diseases and infection. Caulfield, a cardiovascular clinician and researcher, spoke about the UK approach to big data in biomedicine and the role of Genomics England — including how it plans to embed genomic medicine in Britain's National Health Service (NHS).

What are the main challenges to integrating genomic medicine into clinical practice?
The first challenge was to establish a platform that provides the capability and capacity to deliver the programme. To that end, we established 11 genomic-medicine centres across England. These are focused groups of clinicians, scientists and academics that enable us to engage patients, enrol them, receive informed consent, and capture clinical data and samples to analyse.

Another important issue is how to drive up the quality of the interpretation of those genomes. In partnership with the United Kingdom's innovation agency, Innovate UK, we have spent £10 million (US$15.5 million) of government money on stimulating companies to improve the quality of analysis. In December 2014, we instituted a programme called the Genomics England Clinical Interpretation Partnership (GeCIP), which brings together researchers, clinicians and trainees from both the NHS and academia to improve the analysis of genomic data. The GeCIP covers specific domains. For example, we already have one covering haematological oncology, which comprises all the people who work on leukaemia and lymphoma in the United Kingdom.

➲ NATURE.COM
More on the UK 100,000 Genomes Project here:
go.nature.com/ri9rn5

How will your interpretations of the data feed into the health-care system?
If there is an immediately actionable finding, such as a known pathogenic variation in a patient's genome, we send a clinical report directly to the appropriate NHS Genomic Medicine Centre. Clinicians then look at the data and perform their own validation steps to decide whether they think it is correct, before feeding it back to the patient.

But that decision is always with the NHS. This is about creating a genomically enabled community of people who are looking at this data, are familiar with it, and are 'owning' the decision, as they would in the everyday clinical care of those patients. Embedding this autonomy in the NHS will allow us to build a lasting legacy after the initial Genomics England programme has finished.

If we don't find anything that is obviously pathogenic, those genomes go off to the GeCIP domain relating to that patient's illness. This helps to drive up the accuracy of interpreting genomic information concerning the disease. In the meantime, we send reports to the patients, updating them on the progress of the work.

What is the role of industry?
Having a vibrant genomics industry is in the best interests of patients and our community, and of course the total wealth of the country.

We have created a consortium of ten companies ranging from small companies involved in diagnostics or analytics, through to the very large. We have invited those companies into a pre-competitive partnership to look at the first 5,000 genomes with us.

By 'pre-competitive', I mean that they work together, they analyse the data, but they do not own any of the outputs, such as intellectual property. Genomics England owns these on behalf of the UK taxpayer. So if something came up with commercial potential, we would be willing to license that on behalf of the UK taxpayer to third parties, thereby creating the potential for the United Kingdom to draw inward investment in terms of realizing the potential of the resource. This also creates a framework for industry to come in and help shape the programme at the outset.

"The NHS allows us to conjoin academic researchers and the health-care system."

To what extent do you depend on the NHS?
Hospitals and universities in the United Kingdom are all part of one NHS, which allows them to work together cohesively and share information freely — something that would not be possible in a highly competitive and fragmented environment. The NHS is a framework that operates at the level of the whole nation, and it is free at the point of delivery. So it makes a huge difference.

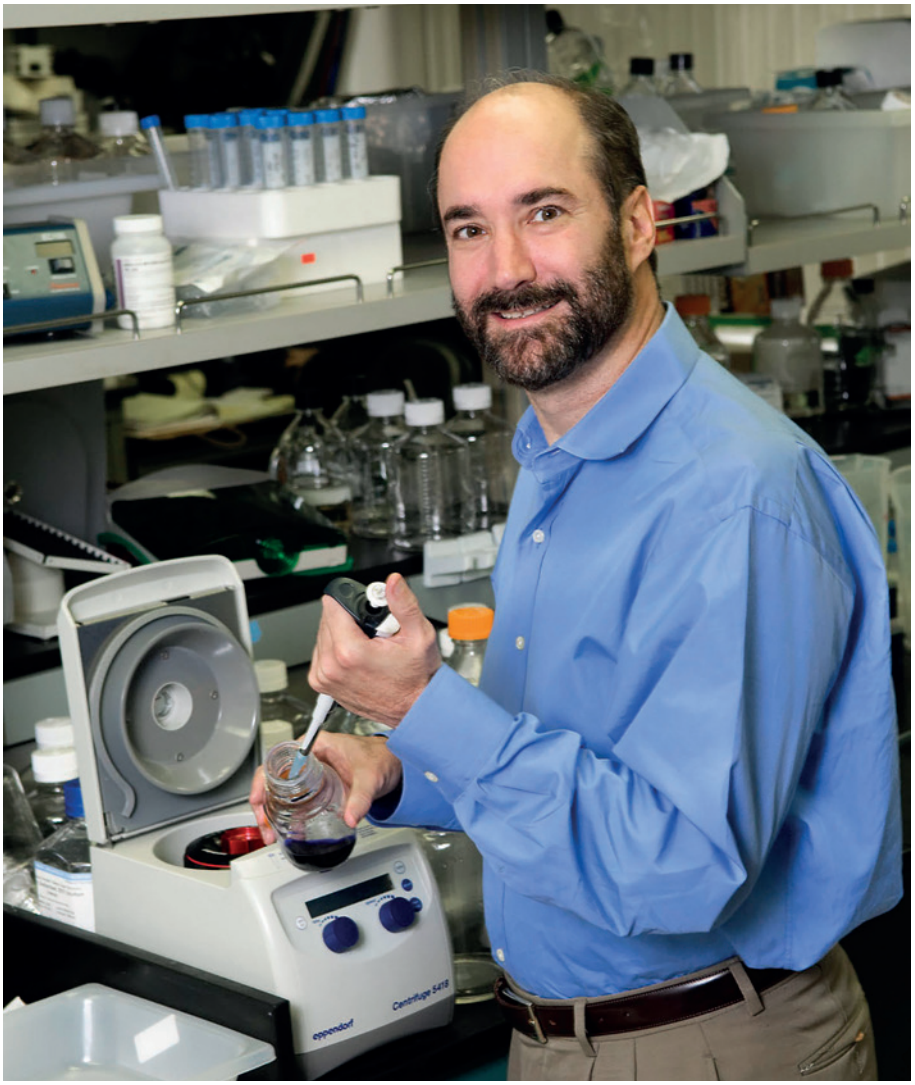
The NHS allows us to conjoin academic researchers and the health-care system so that they can respond rapidly to each other's needs — for example, the health-care system can receive requests to collect data and samples in real time and receive results back quickly.

Do you expect this approach to dramatically speed up research?
It takes an average of 17 years for discoveries to translate from the bench into having a health-care impact. We are seeking to do this in three years. You maximize your opportunity to do that if you juxtapose the health system and the researchers. For people who fund research, this is a hugely effective and efficient way of doing it.

So I see this as a platform not just for a unique transformation of the UK health-care system, but as a model for health-care systems around the world. ■

INTERVIEW BY CLAIRE AINSWORTH

This interview has been edited for length and clarity.



LEE ABEL/STANFORD DEPARTMENT OF GENETICS

Stanford researcher Michael Snyder analysed his own genome, RNA expression and protein production.

PROTEOMICS

High-protein research

The effort to catalogue proteins goes deeper in a push to make genetics research deliver practical benefits.

BY NEIL SAVAGE

When Michael Snyder used the tools of '-omics' on himself, he was in for some surprises. Sequencing his genome, for instance, he discovered that he had a genetic predisposition for type 2 diabetes, even though he did not have any of the standard risk factors, such as obesity or family history of the disease. Over the next 14 months, Snyder, a molecular geneticist at Stanford University in California, repeatedly tested himself to monitor his RNA activity and protein production[1].

When he contracted a respiratory virus midway through the study, he watched as his protein expression changed and biological pathways were activated. Then he was diagnosed with diabetes — it looked to him as if the infection had triggered the condition. He also watched his proteins change during a bout of Lyme disease.

"I had no idea I'd turn out to be interesting," says Snyder, whose body has produced half a petabyte (500,000 gigabytes) of data so far. "It was just a proof of principle."

He has since expanded his study to 100 people, collecting measurements from the proteome and 13 other '-omes', including the proteome and transcriptome of the microorganisms that inhabit their bodies. He hopes that he and others can collect these deep profiles from a million patients, and apply the tools of big data to tease out differences that predict disease and provide a finer-grained understanding of various conditions. He also hopes that they can break conditions down into subtypes by their proteomic profiles. "There are probably 100 different types of diabetes," Snyder says.

Snyder's experience shows the power of using '-omics' to improve our understanding of biology, says William Hancock, a protein chemist at Northeastern University in Boston, Massachusetts.

PRACTICAL GENETICS

Genes provide the instruction manual for biological processes, but it is the proteins they create that turn those instructions into reality. Huge international efforts are under way to identify proteins, map their locations in tissue and cells, count how many are produced in particular circumstances, and describe the various forms they can take. And the oceans of data from these searches will uncover biomarkers for diseases and provide targets for drugs to treat various conditions. By combining proteomics with genomics, transcriptomics, metabolomics and other '-omics', scientists may further deepen their understanding of biology on a molecular level.

Proteomics brings genetic information to a practical level, says Gilbert Omenn, a bioinformatician at the University of Michigan in Ann Arbor and chair of the global Human Proteome Project (HPP). The idea of the project is to create a "complete parts list" of the human body, he says, "to fill in the many blank spots between knowing that a gene has something to do with a disease process and knowing how it really works".

That is quite a parts list. The human body contains roughly 20,000 genes that are capable of producing proteins. Each gene can produce multiple forms of a protein, and these in turn can be decorated with several post-translational modifications: they can have phosphate or methyl groups attached, or be joined to lipids or carbohydrates, all of which affect their function. "The number of potential molecules you can make from one gene is huge," says Bernhard Küster, who studies proteomics at the

NATURE.COM
Find a review of how proteomics affects cell biology here:
go.nature.com/akxnp7

Technical University of Munich in Germany. "It's very hard to estimate, but I wouldn't be surprised to have in one cell type 100,000 or more different proteins."

GLOBAL MAPPING

Proteomics research is an international enterprise. The Human Proteome Organization created two complementary HPP projects, both of which use mass spectrometry. One, the Chromosome-based HPP, divided the 24 chromosomes among 19 countries. Japan, for example, is tackling chromosome 3 and the X chromosome, and Iran is studying the Y. The second, the Biology/Disease-driven HPP, is looking for proteins in specific tissues and organs, focusing on those that are relevant to diseases such as diabetes and colon cancer. A separate global project, the Human Protein Atlas, relies on antibodies with fluorescent molecules or other tags attached that bind to specific proteins to identify them.

There are also some significant national efforts. China is investing heavily in proteomics research, with one example being a new national laboratory called PHOENIX, which was set to open in October with annual funding of US$10 million.

Whatever the technical approach, mapping the human proteome is no easy task. The genome is simple in comparison — it is assembled with just four nucleic acids and changes little over a person's lifetime, except in the special case of cancer. Proteins, on the other hand, vary over time, changing during exercise, disease and menstrual cycles, for example. Another complication is that the most abundant protein can be about 10 billion times as common as the least. "You have one genome and you have a gazillion proteomes, depending on the environmental situation," says Hancock, who is co-chair of the Chromosome-based HPP.

"There is no such thing as a human proteome in one person, let alone in many people," says Küster. Last year, his group published a draft map[2] of a human proteome based on 16,857 mass-spectrometry measurements of human tissue, cell lines and body fluids. They also created a database, ProteomicsDB, to provide analysis of the data.

TOO MUCH DATA?

Just figuring out how to handle the volume of proteomics data is tough. The Human Protein Atlas, for instance, collects images of tissues with tagged antibodies. Each image takes up tens of megabytes, and compressed jpeg files about 10 megabytes in size are made available for online distribution.

Meanwhile, the European Bioinformatics Institute (EBI) in Hinxton, UK, is creating ELIXIR, a distributed-computing infrastructure designed to share proteomics and other biology data among research institutions in Europe. "ELIXIR doesn't want to create a huge database — they want to link different groups and different countries," says Mathias Uhlén, a microbiologist at the KTH Royal Institute of Technology in Stockholm, Sweden. The EBI is already the repository for the Protein Identifications (PRIDE) database, which collects mass-spectrometry data generated by multiple research groups.

But scientists often disagree about whether to keep the raw data or throw it away. "The methods for identifying proteins from raw data are constantly improving, so it makes sense to keep the raw data if you can — but it does take lots of space," says Conrad Bessant, a bioinformatician at Queen Mary University of London. The argument on the other side, he says, is that "the field is advancing so quickly that why would you look at a five-year-old data set? You might as well run the analysis again, because the instruments are so much better."

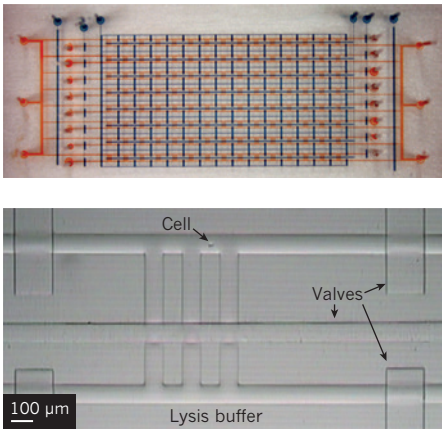
FILLING IN THE MAPS

Proteome data are far from perfect, however. In the issue of *Nature* last May in which Küster's group reported their results, another group of scientists from the United States and India published a draft map[3] said to cover about 84% of the protein-coding genes in the human genome. Both maps were based on mass spectrometry: an enzyme digests proteins and produces peptide sequences about 7 to 30 amino acids long, and the mass of these peptides is used to deduce the protein's composition. And both projects ended up reducing the number of proteins they claimed to have found, after other scientists called into question some of their interpretations[4]. Mass spectrometry is a probabilistic method, says Omenn, and there is no way to exclude the possibility that two different proteins produced the same peptide sequence.

"When it comes to big data, it's easier to generate the data than to get knowledge out of it."

The Human Protein Atlas's antibody-based detection, on the other hand, is non-probabilistic, as it tags individual proteins. The advantage of this approach, argues Uhlén, one of the creators of the Atlas, is that it shows precisely in which organs, tissues and even cells the proteins are located. "What we are providing is a map of where the proteins are," Uhlén says. "That gives you hints about the function of the proteins."

Recent years have seen a push to develop microfluidic chips on which to perform antibody-based single-cell proteomics. This approach is particularly important when the cells of interest are rare, as in the case of circulating tumour cells. It also allows investigators to study differences between populations of the same cell type. For example, if one tumour cell makes many more copies of a particular protein than its neighbour, or the proteins in one cell have a methyl group attached whereas those in another cell do not, this could explain how the tumour develops drug resistance, leading to possible targets for therapeutics.



A proteomics chip (top) profiles individually labelled cells in its microchambers (bottom).

However, even the antibody approach has limitations, as some antibodies can bind to more than one protein, creating misleading results. "An even harder problem is knowing what data are of good quality and what are not," says Uhlén. "When it comes to big data, it's easier to generate the data than to get knowledge out of it."

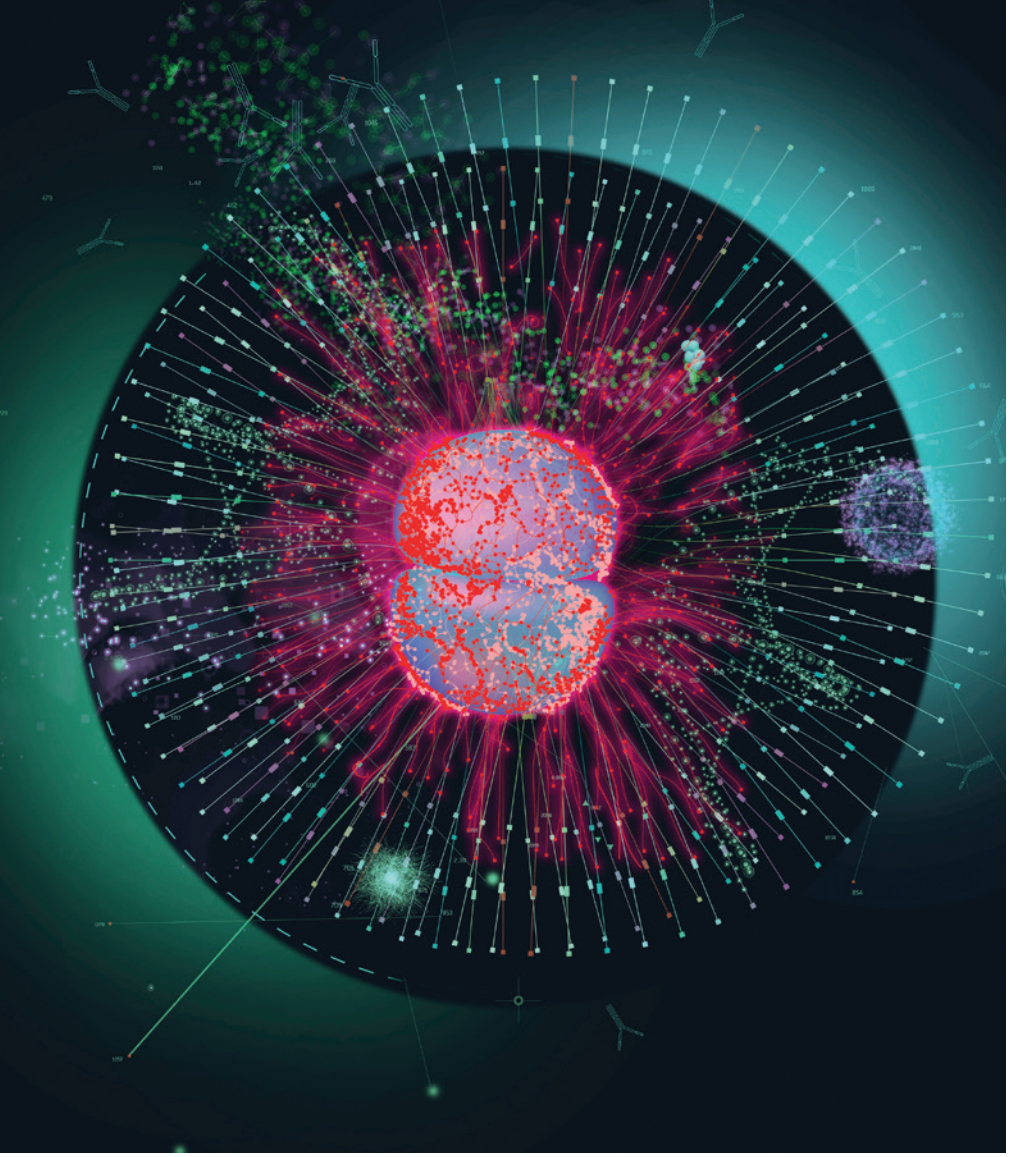
Then there are the missing proteins. Roughly 15% of human genes that should encode proteins have had no associated protein identified[5] — that means there are nearly 3,000 missing proteins. In some cases, this may be because they occur in small amounts or in only tiny areas of tissue. Without a complete catalogue of proteins, the overall picture of human proteomics remains fuzzy.

Computing with incomplete or inaccurate data could lead researchers astray, Hancock worries. "Bringing biology and mathematics together is a match made in hell," he says. "Biology is wet and dirty and messy."

But as measurement techniques improve and scientists amass more findings, "the picture is going to get sharper and sharper," Hancock adds. And the sheer volume of data available to sift through will continue to soar as measurement techniques improve. "We get all kinds of data from many different experiments," Bessant says. "It doesn't take long until you get hundreds of gigabytes or terabytes of data." ■

Neil Savage *is a freelance science and technology writer based in Lowell, Massachusetts.*

1. Chen, R. *et al. Cell* **148,** 1293–1307 (2012).
2. Wilhelm, M. *et al. Nature* **509,** 582–587 (2014).
3. Kim, M.-S. *et al. Nature* **509,** 575–581 (2014).
4. Ezkurdia, I., Vázquez, J., Valencia, A. & Tress, M. *J. Proteome Res.* **13,** 3854–3855 (2014).
5. Horvatovich, P. *et al. J. Proteome Res.* **14,** 3415–3431 (2015).



TATIANA PLAKHOVA

COLLABORATIONS

Mining the motherlodes

Collaboration and competition are spurring on major '–omic' projects.

BY KATHERINE BOURZAC

When actress Angelina Jolie announced in 2013 that she'd had a double mastectomy to reduce her chance of developing breast cancer, after testing positive for a genetic risk factor, the *BRCA* genes responsible were all over the media. These genes carry a significant risk: 55–65% of women with a harmful *BRCA1* mutation, and 45% of women with a mutation in *BRCA2*, develop the disease by the age of 70.

Jolie's case involved a single gene, *BRCA1*, that markedly increased the risk of a specific disease, but the risks of developing genetic diseases are usually much more complicated than that. These complexities are being explored by the many huge research efforts that have been launched in recent years.

Collaborations involving hundreds of scientists and computational biologists are starting to make sense of genomics, proteomics and a host of other '-omics'. Researchers are tracing the twists and turns as thousands of different forms of proteins are churned out and modified (see 'High-protein research', page S6). They are mapping the molecular pathways that flow into or away from different diseases, and are examining the effects of other factors, such as bacteria, on the human body (microbiomics). They are building and testing algorithms to predict how all these '-omic' signatures connect to human health. And they are collaborating to share their ideas and keep each other on track (see 'New eyes on the prize').

➲ NATURE.COM
You can read more about bioinformatics competitions here:
go.nature.com/ihed2h

These large studies make it possible to identify and focus on risk factors for particular diseases. This research, which should enable more personalized treatment for individual patients, is creating huge data sets. Finding rare variations in the genome — and being sure they are not missing something — means sifting through the three billion base pairs in the genomes of tens of thousands of volunteers. To make it work, clinicians from across the world are working with bioinformaticians and computer scientists on a grand scale.

In the process, these researchers also are evolving the art and science of collaboration in the era of big data.

THE QUEST FOR BURIED TREASURE

A disease-focused approach to the genome often involves so-called genome-wide association studies, which are particularly well established in cancer research. In breast cancer, for example, genome-wide association studies have revealed about 90 variants — 'typos' in the genomic code — that are associated with the disease. Of these, only five occur in parts of the genome that code for proteins, says Sara Lindström, a genetic epidemiologist at Harvard University in Boston, Massachusetts.

The other 85 breast-cancer variants are mostly a mystery. "When you see one of these signals, it's not clear if it increases disease risk, or if it's just correlated with disease," says Lindström. Sifting out the important variants requires knowledge of what all these parts of the genome do.

One of the biggest resources for computational biologists tasked with sorting genomic cause from correlation in such puzzles is the Encyclopedia of DNA Elements (ENCODE). Launched in 2003, ENCODE is a mammoth collaborative project funded by the US National Human Genome Research Institute, which maintains a publicly available, searchable genome database.

In 2012, 442 researchers in 32 labs jointly released ENCODE papers that connected more than 80% of the human genome to specific biological functions and identified more than 4 million regions where proteins hook up with DNA (see J. R. Ecker *et al. Nature* **489,** 52–55; 2012 and references therein).

"If you have a favourite gene, you can look it up in ENCODE and find out what regions are likely to regulate that gene," says Michael

Snyder, a Stanford University geneticist and one of the leaders of ENCODE. A breast-cancer researcher, for instance, might find out that a genetic variation uncovered in an association study is a target for a particular transcription factor, a protein that regulates gene expression. That regulatory protein might then be a new target for therapy.

Complementary approaches taken by researchers from 28 institutions are filling in this genome encyclopedia. Many participants study RNA, while some focus on transcription factors or on the regions of the genome where these regulatory elements attach. And still others carry out mapping and data analysis.

Sometimes the sheer size of the ENCODE project can slow things down. A postdoc's idea must be vetted by a larger group, for example, and sometimes researchers have to wait for other labs to finish their work before they can publish a paper, says Manolis Kellis, a computational biologist at the Broad Institute in Cambridge, Massachusetts.

But such problems are far outweighed by the benefits of working together, he says. When you work alone, "bugs can be introduced, and it often takes years to find them", he says. That does not happen in ENCODE — mistakes are usually swiftly spotted by one of a large group of colleagues. The collaborative structure also encourages standardization; researchers need to call a gene or regulatory element by the same name so that they can communicate, and so that the database is searchable and user-friendly.

CANCER IN SEQUENCE

This sort of standardization is essential when dealing with more complex data. The International Cancer Genome Consortium (ICGC), set up in 2008, is trying to deal with this issue at the moment.

The original goal of the project was to sequence the healthy and cancer genomes of 25,000 people. The initial sequencing efforts were performed only on the protein-coding parts of the genome. But consortium leader Tom Hudson, scientific director of the Ontario Institute for Cancer Research in Toronto, Canada, says that now the ICGC has collected about 2 petabytes (2 million gigabytes) of data it plans to go much broader and deeper.

The ICGC will now sequence the non-protein-coding parts of the genome that ENCODE specializes in, and include more clinical information about the patients. This Pan Cancer Analysis of Whole Genomes project will also bring in data from more people — the target is 250,000 — and sequence both their normal and cancer genomes.

This scaling up in the size and scope of the project will be no mean logistical feat. So far the ICGC has brought together leaders from 78 projects in 16 countries. In a pilot of the larger whole-genome comparison project, researchers are analysing paired tumour and normal genomes from 2,600 people, which amounts to about 0.7 petabytes, says Jan Korbel, a computational biologist at the European Molecular Biology Laboratory in Heidelberg, Germany. This is large, but it is still possible to use academic computer centres to process the data.

But the group is at a crossroads. They either need "vast investment" in academic data-centre infrastructure for 250,000 genomes, says Korbel, or they must figure out how to use cloud computing for data sharing and analysis. "You could have several clouds, each specific to a country, as long as those clouds can 'talk' with one another — that is, as long as comparative analyses of data in one cloud with data from another cloud are possible," says Korbel.

"Big data is not particularly useful if you don't have analytics that you can trust."

ANOTHER VIEWPOINT

In efforts like these, standardizing data so that results from different groups are comparable and searchable maximizes the pool of information. This is important when hunting for rare variations that can only be spotted by analysing genomic data from tens of thousands or hundreds of thousands of samples. Working together also helps researchers to strengthen their analyses, says Gustavo Stolovitzky, a computational biologist at IBM's Thomas J. Watson Research Center in Yorktown Heights, New York.

Although big-data analytics can reveal patterns and connections that are otherwise invisible, they can also support a researcher's pre-existing assumptions, thereby obscuring the truth.

One common mistake is 'overfitting'. Stolovitzky likens this to preparing for a university entrance exam by memorizing a big stack of difficult vocabulary flashcards. You can study hard and memorize all the words and their definitions, but that does not mean those words will be on the test — and if they are, the test may use a different wording that throws you off.

Similarly, researchers who devise a predictive algorithm based on their own data set tend to make an algorithm that is good at predicting the results of their own study but fails to work on different data.

Another problem is simply human nature. "When we analyse our own work, we are very benign," says Stolovitzky. It is more useful to involve others, who may have ideas that would never have occurred to someone staring at the same data set all day.

"Big data is not particularly useful if you don't have analytics that you can trust," says Stolovitzky. "We've seen that if you aggregate the results of several algorithms — as long as none of them are bad — the whole is greater than the sum of the parts." That's just one more example of how, when researchers want to get the best results from biomedical big data, working together is crucial. ■

Katherine Bourzac *is a science journalist based in San Francisco, California.*

NEW EYES ON THE PRIZE

Competitions find different ways to solve problems.

Tough problems often benefit from a fresh pair of eyes. That was the thinking in June, when the US National Cancer Institute (NCI) launched a competition called 'Up for a Challenge' to find new ways of analysing breast-cancer data sets. The NCI gathered data from several research groups and is supplying them to teams that present a reasonable proposal, agree to uphold privacy standards, and meet other criteria. The NCI has offered the winner a $30,000 prize and the opportunity to publish a paper in *PLoS Genetics*.

Judges will score entries according to how well groups use innovative methods to find new genetic variants associated with breast cancer, whether the findings can be replicated, and whether they are consistent with known cancer biology. The competition will give extra points to competing groups who formed new collaborations to work on the problem. "We want to reach beyond the usual suspects, and encourage a greater diversity of people to work on these problems," says Elizabeth Gillanders, a genetic epidemiologist at the NCI.

This is one of many competition-based biomedical data projects. Among the others is the DREAM Challenges programme, set up to improve algorithm development in systems biology by Gustavo Stolovitzky, a computational biologist at IBM in Yorktown Heights, New York. The programme has expanded to ask researchers to, for example, predict disease progression and the effectiveness of drug combinations in people with amyotrophic lateral sclerosis.

In many cases, the best performers do not have a background in the specific biology involved. "Presented with a new data set, they shine," says Stolovitzky. K.B.

JARED LAZARUS/FEATURE PHOTO SERVICE FOR IBM

Norman Sharpless of the University of North Carolina works with IBM Watson Health to analyse DNA data.

CANCER

Reshaping the cancer clinic

Big data's war on cancer is still in the early stages, but the front line is advancing.

BY CHARLIE SCHMIDT

The Cancer Genome Atlas, which catalogues cancer mutations, contains some 2.5 million gigabytes of data. This giant project, run by the US National Institutes of Health, has vastly improved our understanding of various forms of cancer — but it holds relatively little information on the clinical experience of the patients who supplied the samples.

At the other end of the cancer treatment chain, electronic health records contain a wealth of case-specific information that could be used to improve cancer care. But more often than not, such records are isolated in individual hospitals and medical practices. As a result, "most patient experiences are lost to research", says Clifford Hudis, an oncologist who specializes in breast cancer at the Memorial Sloan Kettering Cancer Center in New York.

In an effort to improve cancer treatment, Hudis and many others are now collaborating on efforts to bring together and make sense of the big data that emerge from research, patient care and clinical trials. Opportunities for big data extend across most areas of medicine, but "cancer is leading the way", says Lynn Etheredge, a health-care consultant based in Chevy Chase, Maryland. But the ubiquity, variety and lethality of cancer mean that there are plenty of barriers as well as breakthroughs.

NATURE.COM
For more about personalized cancer care, see:
go.nature.com/dktjla

Even so, Etheredge, who in 2007 wrote an influential article for *Health Affairs* calling for "rapid learning systems" to handle big data, believes we have entered a historic period for cancer research and treatment. "We know that cancer is a genetic disease, and we have the databases and the computational power needed to analyse them," he says.

Hoping to build on early successes with personalized cancer drugs, oncologists and computer specialists are working together to harness digitized information and apply it in the clinic. These emerging ventures are competing for business and are grappling with difficult questions about privacy, data ownership and sustainable business models. "Big data is both a research tool and a proprietary commodity," Etheredge says. "It's still early days in the field and there's a lot that we need to work out."

Many organizations and approaches are bringing big data to the cancer clinic in the United States, which leads the world in some aspects of cancer treatment. Here we will consider four: a rapidly growing start-up company, a professional association's initiative, a computer giant's cognitive computing and health-care wing, and a network of academic cancer centres.

THE START-UP

Launched in 2009 by scientists at the Broad Institute in Cambridge, Massachusetts, Foundation Medicine bills insurance companies for its analytical services. Academic and community oncologists submit patients' tissue samples, and Foundation Medicine sequences them. It then screens them for genomic cancer drivers against its own growing database of molecular profiles (generated from more than 50,000 cancer patients so far) and data from other public repositories.

"The public databases aren't like Google — oncologists have no easy way to search them for genomic drivers that relate to their own patient's tumour," says Michael Pellini, chief executive of Foundation Medicine. "So we analyse the tissues and report back available therapeutic interventions, either in the form of a drug approved by the US Food and Drug Administration or a clinical trial."

Oncologists can also query Foundation Medicine's client network for advice on difficult cases. Within 72 hours, Pellini says, responses are aggregated and sent to the doctor, who can then gauge whether a particular drug or approach was effective. The company aims to make its client data more broadly available for use in clinical decision-making.

In January 2015, Swiss pharmaceutical giant Roche spent US$1 billion on a 56% stake in Foundation Medicine, the largest corporate player in this sector, expecting revenue this year of more than $85 million.

PRACTICE MAKES PERFECT

In late 2015, the American Society of Clinical Oncology (ASCO) is expecting to launch CancerLinQ, a platform designed to deliver clinical benefits by analysing aggregated electronic health records from thousands of oncology practices.

Oncologists will be able to interrogate CancerLinQ to see the effects of specific interventions, to review how their own treatment approaches stack up against established care standards, and to develop hypotheses for further study.

"Much of what we know about treating cancer comes from clinical trials that enrol just 3% of the patients diagnosed with cancer every year," says Hudis, who serves on CancerLinQ's board of governors. "With CancerLinQ, we're trying to learn from the remaining 97% who don't participate in these studies."

An initial group of 15 'vanguard practices' of varying sizes are participating in the system, which ASCO expects to contain 500,000 patient records by 2016. Researchers and clinicians will be able to query these records to compare patient outcomes by treatment. Aggregating such large amounts of data should help to reveal the effectiveness of particular drugs or approaches.

"The most important thing that CancerLinQ can do is report on outcomes, for instance, that patients who received a particular treatment lived longer, or had slower progression of their disease," says oncologist Robert Miller, medical director of ASCO's Institute for Quality. These insights will benefit patient care and come at a time, he says, when Medicare, the leading US funder of cancer treatment, is shifting from fee-for-service reimbursement to alternative payment models that reward better outcomes.

A prototype of CancerLinQ was tested in a study of 170,000 breast-cancer patients in 2013. According to Miller, unpublished data showed that the system could highlight trends in data submitted by different medical practices — for example, how they stimulate the production of red blood cells to treat anaemia after chemotherapy.

The platform extracts patient data from electronic health records, anonymizes and aggregates the data, and then integrates them with other types of information, including doctors' notes and biomarker repositories. The goal is eventually to add point-of-care decision support to aid physicians with patients whose diagnosis and treatment is problematic.

CancerLinQ currently relies on donations, but Miller says that in time it will sell effectiveness reports and data-exploration tools to make it more self-sustaining. "We are looking at a range of CancerLinQ-related products and services to help offset the operational costs of the system," says Miller.

COGNITIVE COMPUTING

Big data needs big computing, and in 2013 IBM formed a separate business unit — IBM Watson Health — to focus on commercial opportunities in cancer for its Watson cognitive computing system, which combines natural language and learning capabilities. Watson's store of biomedical knowledge includes every abstract in the PubMed database (there are currently about 25 million and counting); the US National Cancer Institute's Drug Dictionary (which has data on both approved drugs and those in clinical trials); the entire catalogue of somatic cancer mutations in the COSMIC (Catalogue of Somatic Mutations in Cancer) database, which is curated by the Wellcome Trust Sanger Institute, in Cambridge, UK; and data from many other sources.

Watson, which gained fame in 2011 by defeating human champions on the US television quiz show *Jeopardy!*, also has access to anonymized patient data. IBM Watson Health has relationships with more than a dozen medical practices, cancer centres and research organizations, says Ajay Royyuru, director of the Computational Biology Center at IBM Research in Yorktown Heights, New York.

"With CancerLinQ, we're trying to learn from the remaining 97% who don't participate in these studies."

The New York Genome Center relies on Watson to screen DNA mutations in patients enrolled in a study of glioblastoma, an often fatal brain cancer.

Physicians at the Memorial Sloan Kettering centre and at the MD Anderson Cancer Center in Houston, Texas, are training Watson to become a clinical support tool, which entails presenting the computer with anonymized and hypothetical cases. For instance, a patient's tumour might test positive for deficiencies in a gene called *STK11* that may respond to the diabetes drug metformin, Royyuru explains. But Watson might not recommend metformin because this is an off-label indication. "That would be an instance in which it could be taught to cast a wider net," Royyuru says.

Andrew Seidman, a breast-cancer specialist at the Memorial Sloan Kettering centre, adds that the use of Watson must be transparent, so that its reasoning can be easily critiqued. And Seidman cautions that Watson isn't ready for prime time yet. "I'm taking a sober view, and I say that as someone who's helping to develop the technology," he says. In particular, Watson's capacity for natural language processing remains a work in progress. For now, instead of speaking to the computer directly, clinicians have to enter the data manually.

NETWORK NEWS

One of the major challenges facing cancer research is how to match patients with targeted drugs that act on rare mutations, because enrolling enough of these patients in clinical trials is not easy. But one group of hospitals has found a way to get round the problem.

Launched in 2014 by the Moffitt Cancer Center, in Tampa, Florida, the Oncology Research Information Exchange Network (ORIEN) comprises nine academic cancer centres. Patients provide clinical data and tissue samples for analysis, and importantly agree to life-long follow-up, which allows patients to be recruited into new trials geared to their own genetic make-up. "It's a much more proactive way of doing research," says Bill Dalton, ORIEN's founding director.

Moffitt developed the protocol, which it calls "total cancer care", in 2003, and created a company — M2Gen — to handle the analyses and tissue storage in 2006. The development of ORIEN gives this protocol a national reach, with about 130,000 people enrolled so far. Member centres share clinical and molecular data, so they can collaborate on research questions.

BIG PRICE TAGS

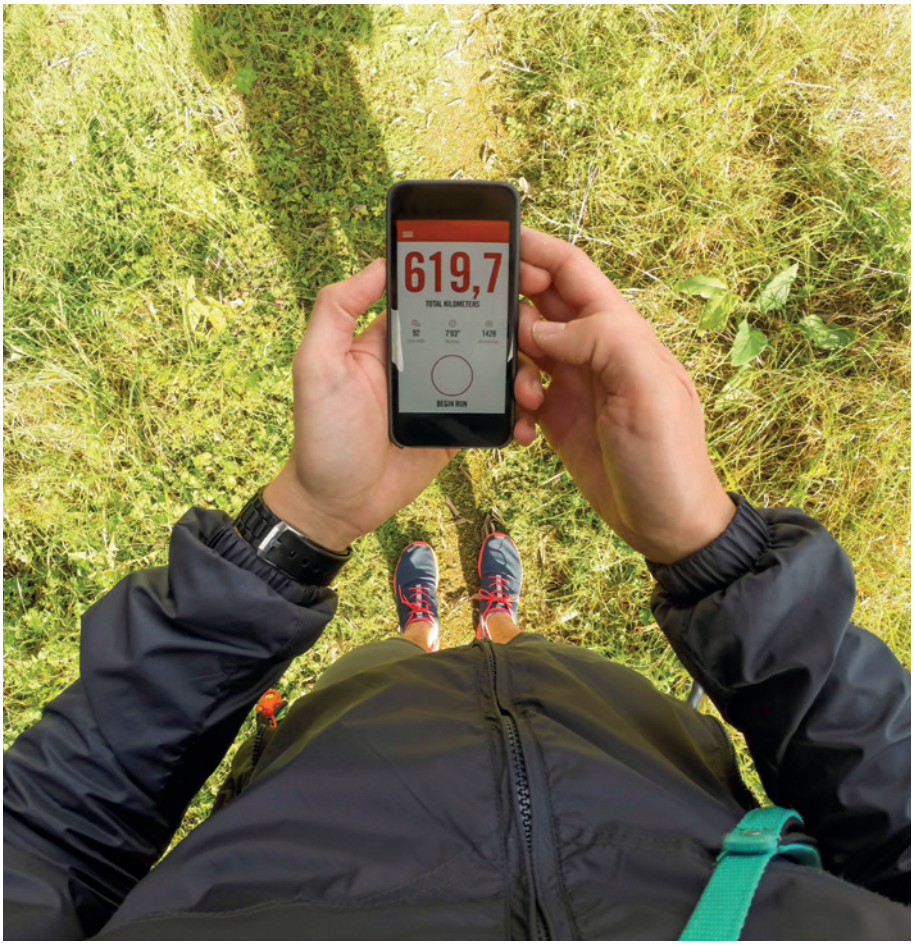
Extracting clinical insights from big data, and using them to guide treatments, does not come cheaply, however. For example, Foundation Medicine charges nearly $6,000 to sequence and interpret the data from a single solid tumour, and more than $7,000 for a blood cancer.

But this is dwarfed by the cost of new oncology drugs, which often have price tags of more than $100,000 per treatment or per year. In July, US Medicare agreed to pay for a leukaemia drug from Amgen that will cost about $178,000 per patient.

Other countries may bargain far more aggressively with drug companies to bring down prices, or reject the drugs altogether on a cost basis, through agencies such as the UK National Institute for Health and Care Excellence.

Ideally, this big money will buy big gains in personalized treatments and cures. This is certainly the hope of the US Medicare and Medicaid officials confronted with spending more than $13 trillion on health care during the coming decade, much of it on cancer therapy. These agencies will wield enormous power over the practicalities of bringing big data into the clinic. Issues relating to data business models and costs will apply across all areas of medicine, "but cancer is forcing them to the table now", says Etheredge. ■

Charlie Schmidt *is a freelance science writer based in Portland, Maine.*



ARTUR DEBAT/GETTY IMAGES

Smartphone fitness apps enable researchers to gather health data from large numbers of people.

MOBILE DATA

Made to measure

Wearable sensors and smartphones are providing a flood of information and empowering population-wide studies.

BY NEIL SAVAGE

For decades, doctors around the world have been using a simple test to measure the cardiovascular health of patients. They ask them to walk on a hard, flat surface and see how much distance they cover in six minutes. This test has been used to predict the survival rates of lung transplant candidates, to measure the progression of muscular dystrophy, and to assess overall cardiovascular fitness.

The walk test has been studied in many trials, but even the biggest rarely top a thousand participants. Yet when Euan Ashley launched a cardiovascular study in March 2015, he collected test results from 6,000 people in the first two weeks. "That's a remarkable number," says Ashley, a geneticist who heads Stanford University's Center for Inherited Cardiovascular Disease. "We're used to dealing with a few hundred patients, if we're lucky."

Numbers on that scale, he hopes, will tell him a lot more about the relationship between physical activity and heart health. The reason they can be achieved is that millions of people now have smartphones and fitness trackers with sensors that can record all sorts of physical activity. Health researchers are studying such devices to figure out what sort of data they can collect, how reliable those data are, and what they might learn when they analyse measurements of all sorts of day-to-day activities from many tens of thousands of people and apply big-data algorithms to the readings.

By July, more than 40,000 people in the United States had signed up to participate in Ashley's study, which uses an iPhone application called MyHeart Counts. He expects the numbers to surge as the app becomes more widely available around the world. The study — designed by scientists, approved by institutional review boards, and requiring informed consent — asks participants to answer questions about their health and risk factors, and to use their phone's motion sensors to collect data about their activities for seven days. They also do a six-minute walk test, and the phone measures the distance they cover. If their own doctors have ordered blood tests, users can enter information such as cholesterol or glucose measurements. Every three months, the app checks back to update their data.

Physicians know that physical activity is a strong predictor of long-term heart health, Ashley says. But it is less clear what kind of activity is best, or whether different groups of people do better with different types of exercise. MyHeart Counts may open a window on such questions. "We can start to look at subgroups and find differences," he says.

It is the volume of the data that makes such studies possible. In traditional studies, there may not be enough data to find statistically significant results for such subgroups. And rare events may not occur in the smaller samples, or may produce a signal so weak that it is lost in statistical noise. Big data can overcome those problems, and if the data set is big enough, small errors can be smoothed out. "You can take pretty noisy data, but if you have enough of it, you can find a signal," Ashley says.

AN APPLE A DAY

Gathering that much data is possible because of Apple software called ResearchKit, which can be used to develop iPhone-based apps for such studies. MyHeart Counts was one of five apps that were launched on the same day that ResearchKit was released. The others are trying to harness the power of big data to study Parkinson's disease, breast cancer, diabetes and asthma.

The Parkinson's study, which enrolled about 16,000 people by July, also uses a walking test, because Parkinson's manifests as a movement disorder. People walk 20 steps in a straight line, and the phone's accelerometer and gyroscope measure their gait to assess their motor control. They are also asked to say "Aaah" for 10 seconds into the phone; measuring how much the voice quavers can help to tell doctors about their muscle tone. "It is very well fitted for using the sensors native to the mobile device," says John Wilbanks, an open-data advocate at Sage Bionetworks, a non-profit biomedical research consultancy based in Seattle, Washington, that developed the Parkinson's mPower app with doctors at the University of Rochester in New York. The app also uses questionnaires

➲ NATURE.COM
Find out more about how mobile tech can aid health research:
go.nature.com/lyvvkp

and can be linked to a fitness tracker to collect even more data.

Similar apps are being written for other smartphone operating systems, such as Windows and Android, and their associated smart watches. There has also been a proliferation of wearable fitness devices from various companies including Basis, Fitbit and Jawbone. Additionally, researchers are developing other types of wearable sensor to collect data over time, including temporary tattoos and contact lenses that measure glucose levels in tears. Meanwhile, existing devices, such as continuous glucose monitors for people with diabetes, are rapidly evolving and adding their data to the mix on smartphones.

Researchers are now trying to use smartphones to go beyond measuring physical fitness. Some, for instance, track mental state and emotional health, by listening to the sound of a person's voice to identify stress, or by tracking their movement to determine their social interaction to figure out if they may be depressed.

As portable devices are increasingly used to measure a whole range of human activity, and computers are now powerful enough to sift through this mountain of data, researchers are hoping to obtain unprecedented insight into human health.

MEASURING UP

The wide variety of measurements from an ever-growing array of devices leaves researchers having to figure out how to handle it all. "It's just an exciting mess," says Ida Sim, co-director of the biomedical informatics division at the University of California, San Francisco.

Sim is a co-founder of Open mHealth, a non-profit company that is developing software to help clean up the mess by standardizing, storing and processing data collected from a variety of devices and apps. "Everybody hates standardization, but without it, it's hard to put data together accurately," she says. For a doctor to correctly interpret a glucose reading, for instance, it is important to know whether that person had been fasting for a period of time.

"You can take pretty noisy data, but if you have enough of it, you can find a signal."

Any effort to establish standards must address two critical questions. How accurate are the readings from these devices? And what exactly is being measured? Today's fitness trackers are designed to tell users whether they have walked more this week than last week, say, not to collect laboratory-quality measurements. "What they know is general movements, which they try to convert to steps, some better than others," says Stephen Intille, who studies personal health informatics at Northeastern University in Boston, Massachusetts.

DIEGO ARGUELLO/STEPHEN INTILLE

Researchers at Northeastern University calibrate sensors during a range of activities in the lab.

To get a better sense of what devices are actually measuring, Intille brings volunteers into his lab and attaches various sensors to both arms and both legs — not only the commercial devices, but other, laboratory-calibrated sensors that record movement, heart rate, breathing and other data points. For 2–3 hours, he takes readings as the volunteers walk, do chores, ride a bicycle and carry out similar activities. Intille then removes some of the sensors and sends the person home, where the remaining devices collect real-world data for another couple of days. For the next three months, he cuts back to the one or two devices he is studying.

This way, Intille can see precisely what the commercial devices are recording during a particular activity. For instance, a Fitbit monitor may produce a certain set of readings when a person is ironing clothes, while the lab equipment records heart rate and breathing. If the computer can be trained to recognize how different activities produce different Fitbit readings in the lab, it may also be able to identify those activities in the real world and analyse their impact on physical fitness.

"I don't personally believe these things are ever going to work really well without some interaction with the end user," says Intille. He wants a phone to tap into data from a fitness tracker and, having learned something about the individual's habits, ask questions, such as: "Are you walking the dog right now?" For a fuller picture, he says, people would need to wear more than one device, perhaps one on the wrist and another on the ankle.

Such detailed information will be needed for researchers to obtain a broader understanding. It's easy to have people report how far and how frequently they run, for example, or how intensely they work out at the gym, but little is known about the effect of day-to-day activity on people's health, says physiologist William Haskell, emeritus professor of medicine at the Stanford Center on Longevity.

"We don't know a lot about the light-intensity range, from standing and just walking about," says Haskell, who has collaborated with Intille and worked to validate the measurements from commercial trackers. "How useful is a standing desk, where you get up and just stand for three hours a day, versus where you have a nice walk around the office? We just don't know."

Haskell started using accelerometers to track physical activity 40 years ago, and he is excited about the possibility of learning from wearable devices. "We think the technology is here," he says. "We just need to validate it and use it to look at a 24-hour activity cycle."

WEAR NEXT?

Obtaining vast amounts of data can improve the power of fitness studies, but wearable technologies also open up the possibility of collecting different kinds of data that were not previously available: the long-term, round-the-clock monitoring of people just going about their business.

"A lot of the promise of big data is that you're not just looking at a lot of data, but you're looking at a lot of data from a lot of different sources," says Sim.

When she sees patients, she interacts with them for about 20 minutes. "For all the time they're not in my clinic, I'm completely blind," she says. "I have no idea what's going on in their lives." Constant data collection could ultimately change that equation and help doctors tailor their care to individual patients. Right now, though, Sim says that there is still a crucial missing link: no one has yet designed a way to send meaningful data from commercial devices to doctors. "It's not built to fit into the physician's workflow at all," she says.

But such pervasive gathering of health information could also offer broader societal benefit. Data from thousands of individuals, collected unobtrusively with technology that is increasingly ubiquitous, could allow for population-wide studies of factors that can affect health. Ashley envisages a mobile-health version of the decades-long Framingham Heart Study, which has helped to identify risk factors for heart disease. He has already started to link the data he is collecting from iPhones with genomic data, which he collects from users who are patients at Stanford Medical Center.

Intille believes that as bigger data sets are created, health researchers will be able to answer a whole range of new questions. "At the individual level we just haven't had any data like this at all," he says. "It's simply not possible to detect it until you have mobile devices. It's totally different from the way we've dealt with health and medicine in the past." ■

Neil Savage *is a freelance science and technology writer based in Lowell, Massachusetts.*

'Outbred' mice are used to reveal the genetic diversity that underlies disease.

CAROLYN A. MCKEONE/SPL

DEEP PHENOTYPING

The details of disease

Precision medicine demands precise matching of deep genomic and phenotypic models — and the deeper you go, the more you know.

BY CATHRYN M. DELUDE

Twenty years ago, amid an explosion of optimism that sequencing the human genome would lead to precision medicine, Isaac Kohane sounded a note of caution. Yes, gene sequencing was a major step forward. But wringing clinical value from the flood of genomic information, he said, would depend on the more pedestrian practice of phenotyping — clinically characterizing traits that signify health or disease, such as a fever, a rash, a limp or an irregular heartbeat.

"Science is informed by what it is possible to measure, and it takes a great leap forward when we can measure something new," says Kohane, a bioinformatician at Harvard Medical School in Boston, Massachusetts. "Previously it was hard to measure differences in genome sequences among individuals. Now that's been reduced to a commodity."

But measuring different phenotypes in diabetes, for instance, still requires someone to comb medical records for data on metrics such as weight, blood pressure and blood glucose levels — a tedious and expensive exercise. Moreover, new forms of measurement, such as continuous glucose monitoring, that may provide valuable clues to disease may not be included in these records.

Precision medicine requires an understanding of the precise relationship between gene and phenotype, and the stratification of diseases into subtypes according to their underlying biological mechanisms. But researchers do not know the functions of most genes, and what they do know is limited to a few cell types, tissues or physiological contexts. Furthermore, descriptions of disease phenotypes often fail to capture the diverse manifestations of common diseases or to define subclasses of those diseases that predict the outcome or response to treatment. Phenotype descriptions are typically "sloppy or imprecise", according to a 2012 review[1].

Overcoming these difficulties requires an exhaustive examination of the discrete components of a phenotype that goes beyond what is typically recorded in medical charts. Such 'deep phenotyping', as it is known, gathers details about disease manifestations in a more individual and finer-grained way, and uses sophisticated algorithms to integrate the resulting wealth of data with other kinds of information.

Historically, phenotyping has not represented big data. It has been partial, generic and time-consuming to gather. Information about individual phenotypes has not been matched to genetic variations among individuals. Deep phenotyping will provide more specificity, new types of big data, and potential connections between disease subtypes and genetic variations.

This approach will allow researchers to address new questions. What is the specific pattern of protein expression or gene regulation in the diseased cells? What about the cells' metabolites and other biochemistry? Are there unusual gut bacteria? Does the patient have other seemingly unrelated conditions, such as autoimmunity or a psychiatric disorder, that might share a biological pathway? This comprehensive deep-phenotyping information, in combination with other big data such as genomic data, can reveal the precise underlying mechanisms of each individual's disease. As Kohane says, deep phenotyping "shows the different dimensions of the disease".

DIVIDING DIABETES

Diabetes exemplifies the problem of imprecise phenotypes. "There are a hundred ways to be diabetic, involving different processes in the pancreas, liver, muscle, brain and fat," says Gary Churchill, a mouse geneticist at the Jackson Laboratory in Bar Harbor, Maine. "Genetic studies lose statistical power by looking at a conglomeration of underlying causes." Different genes are responsible for particular subtypes of diabetes, so mixing them together obscures the reasons why people with the same genetic mutation respond differently to the same treatment.

"There are many steps between causal gene and phenotype at the level of body weight and blood sugar," says Alan Attie, a biochemist at the University of Wisconsin-Madison who collaborates with Churchill. "Each step is subject to genetic variation, which can weaken links between gene and phenotype."

Attie is looking at how individual genomic differences affect one particular phenotype of diabetes: insulin secretion by islet cells. He is isolating islet cells from genetically diverse mice and testing their response not just to

NATURE.COM
Read a feature on population-scale phenotyping here:
go.nature.com/j1yeeg

glucose, but also to fatty acids, amino acids and other molecules that affect insulin secretion. Preliminary data reveal significant variation among islet cells.

Churchill says that studying 'outbred' mice, rather than inbred strains that have identical genomes, better mirrors human diversity in diseases such as diabetes that have many genetic contributors. For instance, B6 mice, a commonly used inbred strain, would all get diabetes when they become obese for the same reason. "If we only studied that mouse, the findings would translate to some human patients but we wouldn't see the breadth of other causes," he says.

BRAIN WORK

Combining deep phenotyping with big '-omic' data is far from straightforward. And the link between gene and phenotype is particularly precarious in neuropsychiatric disorders such as autism.

"Precision medicine? That's not about us. We barely know how to do medicine," says Steven Hyman, a neuroscientist at the Broad Institute in Cambridge, Massachusetts. "In psychiatry, we only have descriptive phenotypes," he says, not mechanistic ones that reveal what has gone awry in the brain. Taking a deep-phenotyping approach to neuropsychiatric disease might break the current impasse in progress to better treatments, says Hyman.

Most brain disorders are polygenic, with different combinations of gene mutations causing disease in individual patients, so identifying genes still fails to explain the majority of cases. For autism, fewer than 10% of cases are linked to genes that might explain the underlying disease mechanism. And an autism gene could also be involved in schizophrenia, obsessive–compulsive disorder and bipolar disorder, says Guoping Feng, a neuroscientist at the Massachusetts Institute of Technology in Cambridge. "Some symptoms are unique to each disorder, but other symptoms overlap."

Furthermore, although most people with autism share core symptoms (such as repetitive behaviours and social deficits), some also have irritable bowel syndrome, infections, seizures, schizophrenia or attention deficit hyperactivity disorder. "We should consider not just neurology and behaviour, but other diagnoses the patient has, such as inflammation and heart disorders," says Kohane. "Defining these subclasses is a prerequisite for precision medicine."

Steve Brown, a mammalian geneticist at the Medical Research Council centre at Harwell, UK, hopes that his work with the International Mouse Phenotyping Consortium can untangle such complications. The consortium is systematically phenotyping a knockout mouse strain for every gene in the mouse genome.

"We can't look at just one or two phenotypes because we don't know the function of most genes," Brown says. "We can't make assumptions about what to look for." Researchers test each mouse for sensory perception, cardiovascular and lung functions, metabolism, morphology and pathologies, and record environmental conditions and diet. They also record behavioural data on activity, social interactions, grooming, sleeping and feeding.

The consortium's knockout mice are all from an inbred strain, which limits the exploration of natural diversity but enables comparative studies and replication of findings. "We never expect to create a model of autism or schizophrenia," Brown says. Instead, the goal is to establish baselines for what each gene does and how it might affect behaviour.

THE LIMITATIONS OF MODELLING

Those who are performing deep phenotyping in animal models acknowledge the fundamental limitations of modelling disorders in non-human species, however. "Human neuropsychiatric disorders involve the prefrontal cerebral cortex, which is a recent arrival in evolution," Hyman says. "Many important cells and circuits in the human cerebral cortex simply aren't there in mice." Scientists should focus on cells and molecular mechanisms that are shared by mice and humans, he says.

"Too many studies start with a transgenic mouse that is, say, lousy at building nests, decide it models schizophrenia or autism, and draw conclusions about the molecular mechanisms of disease," he adds. "It should work the other way round."

Walker Jackson, a prion-disease researcher at the German Center for Neurodegenerative Diseases in Bonn, Germany, studies how single amino-acid mutations in the human prion protein cause the pathologies of Creutzfeldt–Jakob disease and fatal familial insomnia in mice. Jackson measures behaviours to understand the natural history of the diseases, but stops short of seeking a genetic link. "I'm not trying to see how a mutation connects to behaviour because it's hard to know what is changing behaviour," he says.

"Many important cells and circuits in the human cerebral cortex simply aren't there in mice."

He finds that the same mutation affects some neurons but not others, and wants to understand how non-diseased neurons compensate for the mutation to reveal targets for therapy. These effects occur in the hippocampus, cerebellum and thalamus — all regions linked to the behavioural symptoms seen in these disorders. "The data are showing us that the disease is more complex than we thought," Jackson says. "Affected neurons show dysfunction in different ways, so therapy that works in one type of neuron may not work in others."

Similarly, researchers at Stanford University Medical School in California started with a single mutation in the *NL3* gene that has been directly linked to some cases of human autism — a rare occurrence in psychiatric illness. They inserted this mutation in mice and traced its effect on motor behaviour to impaired dopamine inhibition in certain neurons in an unexpected brain region[2].

Feng used a similar approach to identify neural circuitry imbalances caused by another autism gene (*Shank3*) in mice[3]. But this method cannot be widely used because most disorders involve myriad genes, each with a small effect. "I don't think deep phenotyping a mutant mouse's behaviour alone will give us great insight," Feng says. But studying cells derived from humans might help, he suggests, because "these cells already have the perfect combination that can cause disease in a person."

THE HUMAN TOUCH

Given the limitations of animal studies, and the advantages of studying illnesses directly in human cells, deep phenotyping is now extending to research on new human cell models of complex diseases. Neuropsychiatric researchers, for example, can induce skin cells to form stem cells, and can differentiate them into neurons or self-assembled clusters of cells called organoids, so they can study the connections between phenotypes, genomics and related biological data.

Kohane is leading one such project, called N-GRID, which collects cells from patients with neuropsychiatric disorders to look for links between individual genomes and transcriptomes, proteomes, patterns of DNA methylation and other epigenetic markers that affect gene expression, responses to small molecules, and clinical features. The project's deep-phenotyping approach includes "whatever we can measure, to see if distinctive subsets emerge", Kohane says. The aim is to build a "more robust scheme of classifying neuropsychiatric disease — one that is more reliable with regard to prognosis of these diseases, more insightful as to the biological aberration in each category and, therefore, more effective in treating the patient".

Hyman proposes that researchers should consider reserving animal models for safety and pharmacokinetics studies. The efficacy of a new therapy could be tested instead in engineered human cell cultures or organoids. "What if we can't have a mouse model of schizophrenia?" he asks. This should not stop the quest for safe, effective therapies — and if animal models cannot provide good readouts on efficacy, deep phenotyping of human cells might well fill the gaps. ■

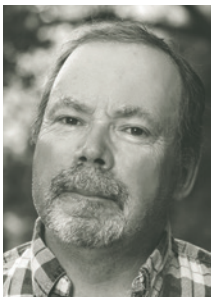
Cathryn M. Delude *is a science writer based in Andover, Massachusetts.*

1. Robinson, P. N. *Hum. Mutat.* **33,** 777–780 (2012).
2. Rothwell, P. E. *et al. Cell* **158,** 198–212 (2014).
3. Peça, J. *et al. Nature* **472,** 437–442 (2011).

PERSPECTIVE

Sustaining the big-data ecosystem

Organizing and accessing biomedical big data will require quite different business models, say **Philip E. Bourne**, **Jon R. Lorsch** *and* **Eric D. Green**.



Biomedical big data offer tremendous potential for making discoveries, but the cost of sustaining these digital assets and the resources needed to make them useful have received relatively little attention. Research budgets are flat or declining in inflation-adjusted terms in many countries (including the United States), and data are being generated at unprecedented rates, so the research community must find more efficient models for storing, organizing and accessing biomedical data. Simply putting more and more money into the current systems is unlikely to work in the long term.

To better understand this situation, we are examining the current and projected costs of managing biomedical data at the US National Institutes of Health (NIH). Our initial analyses indicate that even if we leave out the National Center for Biotechnology Information, which is a special case, the 50 largest NIH-funded data resources have a collective annual budget of US$110 million. And this figure represents just the tip of the iceberg for future needs.

UNDERSTANDING USAGE

Today's biomedical data resources typically treat all items in their collections equally. This does not always make sense, given that the usage patterns of the data vary. But how do we decide which data get more attention? As larger and larger data sets are generated more easily, and the cost of maintaining and annotating these data continues to rise, this question is becoming increasingly important.

Answering it requires a better understanding of how research data are used. This has rarely been thoroughly explored. Historically, funders have been interested primarily in knowing how the data resources that they support are used and by whom. They tended not to look closely at the details of how and why individual items and types of data within a collection are used.

Analyses of these details can be revealing. Preliminary studies suggest that typically a small subset of the data is used frequently, whereas most of the data are rarely accessed. However, the exact subset of data that is used heavily may change over time, and most of the data access may be performed after the data are downloaded, so this is not recorded. All of this means that absolute numbers are hard to interpret.

These caveats notwithstanding, more details of data usage are needed to inform funding decisions. Over time, such usage patterns could tell us how best to target annotation and curation efforts, establish which data should receive the most attention and therefore incur the largest cost, and determine which data should be kept in the longer term. The cost of data regeneration can also influence decisions about keeping data.

Funders should encourage the development of new metrics to ascertain the usage and value of data, and persuade data resources to provide such statistics for all of the data they maintain. We can learn here from the private sector: understanding detailed data usage patterns through data analytics forms the basis of highly successful companies such as Amazon and Netflix.

FAIR AND EFFICIENT

When we have a better understanding of data usage, we can develop business models that consider supply and demand, and develop sustainable practices. In addition, finding economies of scale and harnessing market forces will be essential.

For a typical biomedical data resource, the cost of simply keeping the data is only a small fraction of the total cost of data management. The remainder is largely the cost needed to support the finding, accessing, interoperating and reusing (the FAIR principles; see go.nature.com/axkjiv) of the data — a cost that is widely underappreciated.

Is the FAIR fraction of the cost justified? Are services from different data resources redundant? Are resources subject to 'feature creep' — the addition of costly 'bells and whistles' that are of limited value? Do our funding mechanisms contribute to these problems? And most importantly, is the way we currently maintain biomedical data optimal for the science that needs to be done both today and in the future?

Current practices typically use many disparate sources of data to conduct a study. These data are located in a variety of repositories, often with different modes of access. This lack of centralization and commonality may hinder their ease of use and reduce productivity. We need a better understanding of usage patterns across multiple data resources to use as a basis for redesigning such resources to preserve valuable expertise and curation, and for improving how the data are found, accessed, integrated and reused.

The nature of curation and the quality assurance for biomedical data must also change. Complete and accurate automated or semi-automated extraction of literature is needed to provide metadata and annotation. We should consider crowdsourcing curation, with appropriate validation and incentives. Additionally, the role of professional curators must be better appreciated by data users, by the institutions where the curators work, and by the funders.

THE RESEARCH COMMUNITY MUST FIND MORE **EFFICIENT** MODELS FOR **STORING, ORGANIZING AND ACCESSING** BIOMEDICAL DATA.

In the longer term, we need models that are better aligned with the research life cycle. There is an unnecessary cost in a researcher interpreting data and putting that interpretation into a research paper, only to have a biocurator extract that information from the paper and associate it back with the data. We need tools and rewards that incentivize researchers to submit their data to data resources in ways that maximize both quality and ease of access.

BUSINESS MODELS

One business model worth exploring is the 'freemium model'. Here, the primary data are available free of charge, but services that add value to these data have an associated charge that generates funds that are used to maintain the primary data. This approach is used in other disciplines, notably chemistry. But there are two knotty questions. Should for-profit institutions be charged the same as non-profits? And who should own the intellectual property associated with value-added content?

Another potential business model is the 'subscription model', which is used to access the genetic and molecular databases that are provided by The Arabidopsis Information Resource (TAIR), for example. This option delivers support for a data resource from its active users, but it restricts access, which may be problematic for public-access data policies.

Taking the business-model idea further, what happens if data resources are merged, acquired or go out of business? Would existing resources be more useful and cost-effective if they were merged in some way? Should some services be dropped owing to lack of demand to make way for new services? Would reducing funding for particular data resources over time promote increased efficiency? To answer such questions, we would benefit from advice and help from the private sector and from other scientific communities.

COMMON GROUND

Cloud computing creates an element of data virtualization, takes computing to the data, and may help to solve some of the problems facing biomedical big data. At the NIH, we propose to exploit these opportunities by creating a 'commons' as one possible sustainable model.

Physically speaking, the commons will be collections of public and private resources (including cloud resources) for storing data and computing with those data. To be commons-compliant, such resources must abide by two simple rules. First, each research object in the commons — for example, data, software, narratives or papers — must be uniquely identified, sharable (taking into account privacy issues), and resolvable to its source by using a common identifier. Second, each research object must be defined by a minimal amount of metadata, as defined by the community.

The NIH Big Data to Knowledge (BD2K) programme (bd2k.nih.gov) aims to bring about the creation of the commons. The 12 new BD2K centres are encouraged to share research objects within the commons, and a BD2K consortium is prototyping an index that makes it easy to find commons content.

We also are studying the notion of computing credits, in which a grant recipient is given credits instead of funding to pay for computational time. A principal investigator would be able to spend those credits at any commons-compliant resource. Researchers whose work involves extensive computation on small amounts of data may spend their credits at a different commons-compliant resource to investigators who do minimal computing on large amounts of data.

This model is very different from the situation today. It shifts the initial burden of hardware, data and software maintenance from awardees and their institutions to third parties, notably cloud service providers. The funding model also has the effect of paying only for services used, and aims to create competition in the marketplace, so this approach could result in more data science per dollar.

If the pilot studies at the NIH are successful, it will be important to consider the longer-term implications of a commons model. One outcome is that data and software usage will be tracked both during an award period and after it has expired. Such tracking will yield important usage statistics that can inform future funding decisions.

LEN RUBENSTEIN/THE BROAD INSTITUTE

Research organizations such as the Broad Institute are rapidly evolving their practices for storing and accessing biomedical big data.

UNITING FUNDERS

The medical research community has too little money to start new data resources or to support the growth of more mature databases and services. Moreover, current funding schemes do little to foster the development of best practices; for example, each data resource is usually reviewed in isolation.

Changes to funding practices need to extend across both agency and international borders. Data generation and maintenance are typically funded nationally, but the data are used internationally. As a result, we need to develop more equitable funding models. The first step is for funding agencies to communicate more effectively about data science problems and to seek collaborative solutions. Working from the bottom up, scientists have been doing this for a long time.

Sustaining the biomedical big-data ecosystem is the responsibility of all stakeholders, and will require coordinated efforts among data generators, data maintainers, data users, funders, publishers and others in the private sector. The NIH BD2K programme, in collaboration with many stakeholders, is beginning to address these issues. ■

Philip E. Bourne *is associate director for data science at the US National Institutes of Health. He was previously associate vice-chancellor for innovation and industry alliances at the Office of Research Affairs at the University of California, San Diego.* **Jon R. Lorsch** *is director of the National Institute of General Medical Sciences. He was previously professor of biophysics and biophysical chemistry at Johns Hopkins University in Baltimore, Maryland.* **Eric D. Green** *is director of the National Human Genome Research Institute. He was previously its scientific director, chief of its genome technology branch and director of the NIH Intramural Sequencing Center. e-mail: philip.bourne@nih.gov.*

Q&A Perry Nisen

Better insights, better drugs

A former paediatric oncologist and molecular biologist with experience in academia and industry, Perry Nisen was senior vice-president for science and innovation at GlaxoSmithKline in 2014 before becoming chief executive at the Sanford Burnham Prebys Medical Discovery Institute in La Jolla, California. He discusses the challenges facing drug discovery in the era of big data.

What are some of the biggest obstacles in bringing big data into drug discovery?
Certainly we'll benefit from integrating large data sets, but it is imperative that this is not uncoupled from biological investigation. One of the challenges in pharma, at a time of increased externalization and partnering for research, is how to retain deep biological insight and connect that to the interrogation of these large data sets. One without the other is misguided.

As we move from the lab into the clinic, it is useful to study large, longitudinal clinical data sets. But again, interrogating those data without clinical insight is not very meaningful. The big prizes will go to those who connect the large clinical data sets with an abundance of preclinical data, and bring all that together. I don't think anybody has got that right yet.

In the academic world, the two new drugs to treat high cholesterol that target the protein PCSK9 are a great example of making the connection. Helen Hobbs of the University of Texas Southwestern Medical Center and her partners connected formidable genetics, biological understanding and chemical insight to lead to the important new medicines.

In the pharma world, Genentech seems to have made a long-term investment in biology and linked that to clinical data to treat the right person with the right drug — witness Herceptin for patients with breast cancer.

How are drug companies figuring out where to place their bets?
Given the attention deficit disorder and externalization of research in pharma, and the ever-increasing demands of venture capital and other financial markets in drug discovery, sometimes we are seeing elements of risk aversion and a herd effect. There must be 20 companies looking for the next PD-1 [a cancer immunotherapy target] right now.

"When will we be comfortable enough to invest substantially in making medicines to alter someone's microbiome?"

But we are also seeing examples of pharma companies making very big, bold decisions — going after certain bespoke immune therapies, for example, without any hugely compelling evidence that this approach will work.

I think there is an argument for companies externalizing and partnering on research, so that they are not missing something at the bleeding edge of discovery. At the same time, those decisions of when and where you jump in, with enough confidence, are truly challenging.

What is a good example of a tough decision?
Look at microbiome research. Everyone is really excited about the microbiome. It is tantalizing that all these billions of bacteria in our gut, skin and everywhere else influence disease and how we respond to drugs. We have seen associations between particular bacterial flora and disease states, but when will we be comfortable enough to invest substantially in making medicines to alter someone's microbiome? I don't know the answer to that.

To date, most of the data set out to define the diverse repertoire of microbes are from small numbers of individuals. Conducting evaluations from larger cohorts of subjects is daunting, and long-term clinical data on these cohorts have been limited.

Finally, we lack the robust tests that would let us modify the microbiome in a durable way and have a meaningful impact on the disease process.

Why does Sanford Burnham Prebys combine basic research with drug discovery?
Pharma often has a disconnect between the applied research in making medicines and the deep biological insight and ongoing experimentation that informs it. We have 80 people here from pharma who have made making medicines their whole career. Our principal investigators can work hand-in-hand with drug discoverers, and that enables us to pursue research and go after targets that nobody else will work on because they are too uncertain.

Building up the big-data element creates a unique situation in this work. For example, when we start thinking about autoimmunity, about what happens when, with which T cells and B cells, how do we start disentangling those complexities?

This is where big-data bioinformatics can be hugely useful. That to me underscores more than ever the need both to analyse large data sets and to stay connected to researchers who understand all the moving pieces with a view from a little higher up.

Is it difficult to find research staff with skills in gathering and analysing big data?
Yes, we struggle with this. Part of the challenge is training and funding individuals who are sophisticated enough to pursue that. They also tend to be gobbled up by potentially more lucrative fields outside the life sciences.

We have been searching hard to attract and recruit the next wave of systems biologists and other people who can analyse large amounts of data. You want to have a critical mass of people doing that together, and it's really tough. There are people who generate data and people who analyse data, but there are few who do both really well. I believe the winners are the ones who can pull it all together. ■

INTERVIEW BY ERIC BENDER

This interview has been edited for length and clarity.

TATIANA PLAKHOVA

BIG DATA IN BIOMEDICINE

4 BIG QUESTIONS

Gathering and understanding the deluge of biomedical research and health data poses huge challenges. But this work is rapidly changing the face of medicine.

BY ERIC BENDER

	QUESTION	WHY IT MATTERS	NEXT STEPS	QUOTE
1	***How can long-term access to biomedical data that are vital for research be improved?***	Data storage may be getting cheaper, particularly in cloud computing, but the total costs of maintaining biomedical data are too high and climbing rapidly. Current models for handling these tasks are only stopgaps.	Researchers, funders and others need to analyse data usage and look at alternative models, such as 'data commons', for providing access to curated data in the long term. Funders also need to incorporate resources for doing this.	"Our mission is to use data science to foster an open digital ecosystem that will accelerate efficient, cost-effective biomedical research to enhance health, lengthen life and reduce illness and disability." **Philip Bourne**, US National Institutes of Health.
2	***How can the barriers to using clinical trial results and patients' health records for research be lowered?***	'De-identified' data from clinical trials and patients' medical records offer opportunities for research, but the legal and technical obstacles are immense. Clinical study data are rarely shared, and medical records are walled off by privacy and security regulations and by legal concerns.	Patient advocates are lobbying for access to their own health data, including genomic information. The European Medicines Agency is publishing clinical reports submitted as part of drug applications. And initiatives such as CancerLinQ are gathering de-identified patient data.	"There's a lot of genetic information that no one understands yet, so is it okay or safe or right to put that in the hands of a patient? The flip side is: it's my information — if I want it, I should get it." **Megan O'Boyle**, Phelan-McDermid Syndrome Foundation.
3	***How can knowledge from big data be brought into point-of-care health-care delivery?***	Delivering precision medicine will immensely broaden the scope of electronic health records. This massive shift in health care will be complicated by the introduction of new therapies, requiring ongoing education for clinicians who need detailed information to make clinical decisions.	Health systems are trying to bring up-to-date treatments to clinics and build 'health-care learning systems' that integrate with electronic health records. For instance, the CancerLinQ project provides recommendations for patients with cancer whose treatment is hard to optimize.	"Developing a standard interface for innovators to access the information in electronic health records will connect the point of care to big data and the full power of the web, spawning an 'app store' for health." **Kenneth Mandl**, Harvard Medical School.
4	***Can academia create better career tracks for bioinformaticians?***	The lack of attractive career paths in bioinformatics has led to a shortage of scientists that have both strong statistical skills and biological understanding. The loss of data scientists to other fields is slowing the pace of medical advances.	Research institutions will take steps, including setting up formal career tracks, to reward bioinformaticians who take on multidisciplinary collaborations. Funders will find ways to better evaluate contributions from bioinformaticians.	"Perhaps the most promising product of big data, that labs will be able to explore countless and unimagined hypotheses, will be stymied if we lack the bioinformaticians that can make this happen." **Jeffrey Chang**, University of Texas.

Eric Bender *is a freelance science writer based in Newton, Massachusetts.*